

e-ISSN 2791-7835



# LOKMAN HEKIM HEALTH SCIENCES

Volume 6 Issue 2 Year 2026

[www.lokmanhekimhs.com](http://www.lokmanhekimhs.com)





**EDITOR-IN-CHIEF****Prof. Mehmet DOĞANAY, MD**

Department of Infectious Diseases, Lokman Hekim University Faculty of Medicine, Ankara, Türkiye

**ASSOCIATE EDITOR-IN-CHIEF****Assoc. Prof. Hülya ŞİRİN**

Department of Public Health, Gülhane Faculty of Medicine, University of Health Sciences, Ankara, Türkiye

**ASSOCIATE EDITORS****Assistant Prof. Açelya ERİKÇİ**

Department of Biochemistry, Lokman Hekim University, Faculty of Pharmacy, Ankara, Türkiye

**Assoc. Prof. Zeynep TÜRE YÜCE**

Department of Infectious Diseases and Clinical Microbiology, Afyonkarahisar State Hospital, Afyonkarahisar, Türkiye

**FIELD EDITORS****Medical Sciences****Prof. Emine Elif ALTUNTAŞ**

Department of Otolaryngology, Lokman Hekim University Faculty of Medicine, Ankara, Türkiye

**Assoc. Prof. Mehmet AZİRET**

Gastroenterology Surgery Unit, Department of Surgery, University of Health Sciences, Balıkesir City Hospital Health Application and Research Center, Balıkesir, Türkiye

**Prof. Gökçen DİNÇ**

Department of Microbiology, Erciyes University Faculty of Medicine, Kayseri, Türkiye

**Assoc. Prof. Dr. Suna KARACA**

Department of Obstetrics and Gynecology, Faculty of Medicine, İzmir Bakırçay University, İzmir, Türkiye

**Assoc. Prof. Murat Bülent KÜÇÜKAY**

Department of Internal Medicine, University of Health Sciences, Gülhane Faculty of Medicine, Dr. Abdurrahman Yurtaslan Oncology Hospital, Ankara, Türkiye

**Prof. Deniz ODABAŞ**

Department of Public Health, Faculty of Medicine, Ankara University, Ankara, Türkiye

**Assoc. Prof. Aydan ÖRSÇELİK**

Department of Sports Medicine, Gülhane Faculty of Medicine, University of Health Sciences, Ankara, Türkiye

**Prof. Apar PATAER**

Department of Thoracic and Cardiovascular Surgery - Research, Division of Surgery, The University of Texas MD Anderson Cancer Center, Houston, Texas, United State of America

**Assoc. Prof. İsmail SARI**

Department of Medical Biochemistry, Faculty of Medicine, Kırklareli University, Kırklareli, Türkiye

**Prof. Hasan TEZER**

Department of Pediatrics, Gazi University Faculty of Medicine, Ankara, Türkiye

**Prof. Levent YAMANEL**

Medical Intensive Care Unit-Department of Internal Medicine, Gülhane Faculty of Medicine, University of Health Sciences, Ankara, Türkiye

**Pharmaceutical Sciences****Prof. Gökalp İŞCAN**

Department of Pharmacognosy, Anadolu University Faculty of Pharmacy, Eskişehir, Türkiye

**Assoc. Prof. Nilüfer ORHAN**

American Botanical Council, ABC AHP NCNPR Botanical Adulterants Prevention Program, Austin, USA

**Prof. M. Betül YERER AYCAN**

Department of Pharmacology, Erciyes University Faculty of Pharmacy, Kayseri, Türkiye

**Dental Sciences****Prof. Kadriye Funda AKALTAN**

Ankara University Faculty of Dentistry, Ankara, Türkiye

**Assoc. Prof. Banu SAĞLAM AYDINATAY**

Department of Orthodontics, Lokman Hekim University Faculty of Dentistry, Ankara, Türkiye

**Assoc. Prof. Zeynep BURÇİN GÖNEN**

Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Erciyes University, Kayseri, Türkiye

**Assoc. Prof. Katibe Tuğçe TEMUR**

Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Ömer Halisdemir University, Niğde, Türkiye

**Assoc. Prof. Samet TOSUN**

Department of Endodontics, Faculty of Dentistry, Pamukkale University, Denizli, Türkiye

**Prof. Burak YILMAZ**

School of Dental Medicine, Bern, Bern University, Switzerland and Ohio State University, Ohio, USA

**Health Sciences****Prof. Nilgün BEK**

Department of Physiotherapy and Rehabilitation, Lokman Hekim University Faculty of Health Sciences, Ankara, Türkiye

**Prof. Gülten GÜVENÇ**

Department of Obstetrics and Gynecology Nursing, Gulhane Faculty of Nursing, University of Health Sciences, Ankara, Türkiye

**Dr. Monika Gostic, PhD**

Department of Medical Sciences and Nutrition, School for Medicine, Aberdeen University, Aberdeen, United Kingdom

**Assoc. Prof. Ela HIÇYORULMAZ**

Department of Health Systems, Faculty of Health Sciences, Hitit University Çorum, Türkiye

**Prof. Figen İNCİ**

Department of Psychiatry, Faculty of Health Sciences, Ömer Halisdemir University, Niğde, Türkiye

**Assoc. Prof. Özgü İNAL ÖZÜN, PhD**

Department of Neurologic Physiotherapy Rehabilitation, Gülhane Faculty of Physiotherapy and Rehabilitation, University of Health Sciences, Ankara, Türkiye

**Assoc. Prof. Dr. Hale SEZER**

Department of Nursing Education, Faculty of Health Science, İzmir Bakırçay University, İzmir, Türkiye

**Assistant Professor Habiba SHAH, PhD**

Department of Nutrition &amp; Food Science, Dow University of Health Sciences, Faculty of Public Health, Karachi, Pakistan

**PUBLISHER****Kare Media** / Publication Coordinator: *Ece Hanne Şimşek* / Graphics, Design: *Neslihan Çakır***Address:** Göztepe Mahallesi, Fahrettin Kerim Gökay Caddesi, No: 200, Daire: 2, Göztepe, Kadıköy, İstanbul, Türkiye **Phone:** +90 216 550 61 11 / **Fax:** +90 216 550 61 12 / **E-mail:** kare@karepb.com

**LINGUISTIC EDITOR****Assistant Prof. Gohar SEYEDİ**Director of Language Teaching Practice and Research Center, Rectorate, Lokman Hekim University,  
Ankara, Türkiye**BIostatISTICS EDITOR****Müge COŞKUN YILDIRIM**Department of Biostatistics, Lokman Hekim University Faculty of  
Medicine, Ankara, Türkiye**Prof. Ergün KARAAĞAOĞLU**Department of Biostatistics, Lokman Hekim University Faculty of  
Medicine, Ankara, Türkiye**ADVISORY BOARD****Prof. Hasan ALKUMRU**Department of Dentistry, University of  
Western Ontario, Canada**Prof. Selen ÖZAKAR AKÇA**Department of Child Health and Diseases  
Nursing, Faculty of Health Sciences Hitit  
University Çorum, Türkiye**Prof. Les BAILLIE**School of Pharmacy and Pharmaceutical  
Sciences Microbiology, Ecosystem and  
Biodiversity, Antibacterial Compounds,  
Cardiff University, Cardiff,  
Wales, UK**Prof. Sarah A. CURRAN**School of Sport and Health Sciences,  
Cardiff Metropolitan University, Western  
Avenue, UK**Prof. İlhami ÇELİK**Department of Infectious Diseases,  
University of Health Sciences, Faculty of  
Medicine, Kayseri, Türkiye**Prof. Metin DALIP**

University of Tetova, Tetova, Macedonia

**Prof. Engin DURSUN**Faculty of Medicine, Lokman Hekim  
University, Ankara, Türkiye**Prof. Jianbo XIAO**Department of Analytical Chemistry and Food  
Science, University of Vigo, Vigo, Spain**Prof. Saule MAUKEYAVA**Department of Infectious Diseases, Semey  
Medical University, Faculty of Medicine,  
Semey, Kazakhstan**Prof. Bülent ÖZPOLAT**Department of Nanomedicine, Houston  
Methodist Research Institute, USA**Prof. Elizabeth C. POSTER**

University of Texas, Arlington, USA

**Chiara SINIGAGLIA**Department of Science, University of  
Basilicata, Potenza, Italy**Prof. Jafar SOLTANI**Department of Pediatrics, Faculty of  
Medicine, Tehran University of Medical  
Sciences, Tehran, Iran**Prof. Sultan TAŞÇI**Department of Internal Medicine Nursing,  
Erciyes University, Faculty of Health  
Sciences, Kayseri, Türkiye**Prof. Burak YILMAZ**School of Dental Medicine, Bern University,  
Bern, Switzerland and Ohio State University,  
Ohio, USA**PREVIOUS EDITOR-IN-CHIEFS****Prof. Feyyaz ONUR (2021–2024)****Owner (Behalf of Lokman Hekim University)****Mehmet ALTUĞ, MD****Holder of a Concession (Behalf of Lokman Hekim University)****Prof. Fatih GÜLTEKİN, MD****ABSTRACTING AND INDEXING**Lokman Hekim Health Sciences is indexed in **TRDizin, EBSCO, CINAHL, Open Ukrainian Citation Index (OUCI), Scope Database, Scilit, Google Scholar, DOAJ, Gale, Research4life, Hinari, IdealOnline, MIAR, TürkMedline, ASCI, WorldCat, OpenAlex, EZB, EuroPub and Embase**

## INSTRUCTIONS FOR THE AUTHORS

### AIM&SCOPE

Lokman Hekim Health Sciences is an open access journal that publishes national and international manuscripts in the field of health sciences. The journal accepts original articles, review / meta-analysis, brief report, case report, letter to the editor, and manuscripts related to the history of medicine in different disciplines in the field of health sciences.

All manuscripts must be submitted electronically to <https://lhhs.manuscriptmanager.net/>.

Lokman Hekim Health Sciences is a journal that publishes manuscripts in the field of medicine, pharmacy, dentistry, sports and health sciences (nursing, physical therapy and rehabilitation, nutrition and dietetics, midwifery, ergotherapy, audiology, speech and language therapy).

- The language of the journal is English
- It is an international electronic journal.
- Publishes scientific articles in general health sciences by evaluating them within the framework of the principles of refereeing.
- Lokman Hekim Health Sciences is an open access journal; that is published articles can be accessed on the journal's online page without any subscription or fee. LHU Open Access Policy is taken as basis for the journal to be open access.
- Article submission and access is free.
- The journal is published 4 issues in a year.

### EDITORIAL POLICY

The editorial and publication processes of the journal are shaped in accordance with the international guidelines. Original manuscripts that have not been published before in an electronic or printed medium in any language, having high specific quality and citation potential will be accepted for publication. Manuscripts that have been presented in a meeting before should be submitted with detailed information of the event. Any question related to the editorial process should be forwarded to the secretariat, managing editor or the editor-in-chief and should be sent to [kare@karepb.com](mailto:kare@karepb.com).

All contents are the responsibility of authors. All financial liability and legal responsibility associated with the copyright of submitted tables, figures and other visual materials protected by national and international laws rest with the authors. The authors take responsibility for any legal proceedings issued against the journal.

### REVIEW PROCESS

All submitted manuscripts will undergo a double-blind peer-review process. At least two independent reviewers who are experts in their field will evaluate the publication potential of the manuscripts within four weeks. The editor-in-chief has the final decision-making opportunity for all submissions. For the evaluation process of manuscripts submitted by the editorial board members of the journal, an external and independent editor will be invited.

If any revision is advised, the revised manuscript should precisely indicate every step taken in accordance with the reviewers' notes.

### ETHICS

Studies using human or animal subjects should be approved by the appropriate institutional and local Ministry of Health ethics committees. Ethics approval of research protocols in accordance with international agreements (World Medical Association Declaration of Helsinki "Ethical Principles for Medical Research Involving Human Subjects," amended in October 2013, [www.wma.net](http://www.wma.net)) is required for experimental, clinical, and drug studies, as well as for some case reports. Ethics committee reports or an equivalent official document may be requested from the authors.

In all manuscripts, a statement regarding the patient consent, and the details of the ethical approval (name of the ethics committee, date and number) should be stated in the Materials and Methods section. For photographs that may reveal the identity of the patients, signed releases of the patient or of their legal representative should be enclosed.

### AUTHORSHIP

Each individual listed as an author should fulfill the authorship criteria recommended by the International Committee of Medical Journal Editors (ICMJE - [www.icmje.org](http://www.icmje.org)).

The ICMJE recommends that authorship be based on the following 4 criteria:

1. Substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work; AND
2. Drafting the work or revising it critically for important intellectual content; AND
3. Final approval of the version to be published; AND
4. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

For the manuscript submission, it is the responsibility of corresponding authors to submit a signed and scanned version of the authorship contribution form (available for download through <https://lokmanhekimhs.com/assets/pdfs/author-contribution-form.pdf>). All authors and the order of authors should be listed carefully at the time of the original submission. Any addition, removal or re-arrangement author names in the authorship list should be made only if approved by the journal Editor before the manuscript has been accepted. For such an alteration, corresponding author should define the reason for the change in author list in detail with a written confirmation from all authors that they agree this alteration.

### ORCID ID

The Open Researcher and Contributor ID (ORCID) number of each author must be submitted when creating an account for correspondence. To obtain an ORCID number, please visit <https://orcid.org/>

### DISCLOSURE AND CONFLICTS OF INTEREST

The Lokman Hekim Health Sciences requires the authors of submitted manuscripts to disclose any existing or potential conflicts of interests, including financial, consultant, and institutional relationships, that might lead to potential bias or a conflict of interest. The ICMJE Potential Conflict of Interest Disclosure Form should be completed and submitted by all contributing authors to disclose any potential conflict of interest. Cases of a potential conflict of interest of the editors, authors, or reviewers are resolved by the journal's editorial board according to COPE and ICMJE guidelines.

The Lokman Hekim Health Sciences requires each submission to be accompanied by a Copyright Transfer Form (available for download at <https://lokmanhekimhs.com/assets/pdfs/copyright-transfer-form.pdf>) assigned by all authors. When using any material previously published, the authors must obtain permission from the copyright holder. Legal, financial, and criminal liabilities in this regard belong to the author(s).

### PLAGIARISM DETECTION

All submissions are screened using similarity detection software (iThenticate; Turnitin LLC, Oakland, CA, USA) at least two times: on submission and before publication. In the event of alleged or suspected research misconduct, the editorial board will follow and act in accordance with COPE guidelines.

### OPEN ACCESS AND PUBLICATION CHARGES

Lokman Hekim Health Sciences is an open access journal, which means that all content is freely available to read, download, copy, distribute, print, search, or link to the full texts of the articles, without charge to any user. There are no submission fees, publication fees, or page charges for this journal.

### ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

At submission, the journal should require authors to disclose whether they used artificial intelligence (AI)-assisted technologies (such as Large Language Models [LLMs], chatbots, or image creators) in the production of submitted work. Authors who use such technology should describe, in both the cover letter and the submitted work, how they used it. Use of AI for writing assistance should be reported in the acknowledgment section. Authors who used AI technology to conduct the study should describe its use in the methods section in sufficient detail to enable replication to the approach, including the tool used, version, and prompts where applicable. Chatbots (such as ChatGPT) should not be listed as authors because they cannot be responsible for the accuracy, integrity, and originality of the work, and these responsibilities are required for authorship. Therefore, humans are responsible for any submitted material that included the use of AI-assisted technologies. Authors should carefully review and edit the result because AI can generate authoritative-sounding output that can be incorrect, incomplete, or biased. Authors should not list AI and AI-assisted technologies as an author or co-author, nor cite AI as an author. Authors should be able to assert that there is no plagiarism in their paper, including in text and images produced by the AI. Humans must ensure there is appropriate attribution of all quoted material, including full citations.

### MANUSCRIPT PREPARATION

Manuscripts should be prepared in accordance with the international guidelines.

Manuscripts may only be submitted through the journal's online manuscript submission and evaluation system, <https://lhhs.manuscriptmanager.net/>.

The first step in the evaluation is the technical evaluation process. In the technical evaluation, the editorial staff will ensure that the manuscript has been prepared and submitted in accordance with the journal's guidelines. Manuscripts that do not conform to the journal's guidelines will be returned to the author with requests for technical correction. The quality and clarity of the language used in a manuscript is very important. Please refer to specific formatting requirements noted in the submission checklist, not to cause any delays for your manuscript.

### PREPRINT POLICY

Lokman Hekim Health Sciences welcomes the submission of manuscripts that have been previously deposited on a recognized non-commercial preprint server. Preprints are not considered prior publications, allowing authors to disseminate and discuss their findings with the scientific community before formal peer review and publication in the journal. Authors are required to disclose the details of the preprint during the initial submission process. This includes providing the journal with the deposition information, such as the name of the preprint server and the DOI or any other unique identifier associated with the preprint. Once the article is accepted and published in the journal, authors are responsible for updating the preprint record to include a link to the final published version. This ensures proper citation and access to the most reliable and peer-reviewed version of the research. This policy reflects the journal's commitment to promoting transparency, accessibility, and the dissemination of scientific knowledge.

### MANUSCRIPT FORMATTING AND TYPES

The manuscript should be typed in a Microsoft Word™ file, single-column format, double-spaced with 2.5 cm margins on each side, and 12-point type in Times New Roman font.

All abbreviations in the text must be defined the first time they are used (both in the abstract and the main text), and the abbreviations should be displayed in parentheses after the definition. Authors should avoid abbreviations in the title. Measurements should be reported using the metric system according to the International System of Units (SI). When a drug, product, hardware, or software mentioned within the main text product information, including the name of the product, producer of the product, city of the company and the country of the company should be provided in parenthesis.

**Editorial comment:** Editorial comments provide a brief critical commentary by an invited experienced author in the topic of a research article previously published in the journal. The word count is limited to 1200 and 10 references may be included.

The submission should not include an abstract, keywords, tables, figures, and images.

**Case report:** Reports of rare cases or conditions that reflect challenges in diagnosis and treatment, or present something otherwise particularly interesting and educative will be accepted. It should contain an unstructured abstract of a maximum of 150 words and the text should

be structured with subheadings of introduction, case report, and discussion. A case report is limited to 1200 words and 10 references.

**Brief report:** A brief report expresses a focused message in an interesting topic. It can include small case series, negative trials, the preliminary results and others that are not to be published as a full text paper. Brief report is restricted to a maximum of 1500 words, no more than 1 table and 3 figures, and 15 references. It should contain an unstructured abstract of a maximum of 150 words.

**Original article:** It provides new information based on an original and novel research. It should contain a structured abstract of a maximum of 250 words with the following subheadings: Background and Aims, Methods, Results, and Conclusion. The main text of an original article should be structured with Introduction, Methods, Results, Discussion, Conclusion, Acknowledgments, References, Tables, and Figure Legends subheadings. Original articles are limited to 3500 words and 30 references.

**Review article:** There are two types of review articles; Narrative reviews and Systematic reviews.

For the narrative reviews, the authors who have extensive knowledge and a scientific background of a particular field will be invited by

the journal staff. This type of reviews evaluates the current level of knowledge of a topic in clinical practice and should guide future studies. All invited review articles will also undergo peer review prior to acceptance. The main text should contain introduction, clinical and research consequences, and conclusion sections. At the end of the introduction, a short description of how the articles reviewed have been selected (keywords, time period of the search, etc.) should be included.

In systematic reviews, a direct question should be addressed. Data sources, study eligibility criteria, participants, interventions, and statistical analysis should be defined in detail. The main text of a systematic review should contain Introduction, Methods, Results, Discussion, and Conclusion subheadings.

**History of medicine:** The manuscripts related to the history of medicine, prepared by authors who have extensive knowledge of a particular historical field, invention, surgical procedure, or application are also welcomed. The text should include an unstructured abstract of a maximum 150 words, keywords, a short introduction, historical background of the subject, and a conclusion. This type of manuscript should not exceed 1500 words.

---

**Authors are required to prepare manuscripts in accordance with the international guidelines\* below**

Randomized Controlled Trial**	<b>CONSORT</b> (Consolidated Standards of Reporting Trials)
Non-Randomized Trial**	<b>TREND</b> (Transparent Reporting of Evaluations with Non-randomised Designs) ( <a href="https://www.cdc.gov/trendstatement/index.html">https://www.cdc.gov/trendstatement/index.html</a> )
Trial Protocol	<b>SPIRIT</b> (Standard Protocol Items Recommendations for Interventional Trials)
Observational Epidemiologic Studies (cohort, case-control, cross-sectional)	<b>STROBE</b> (the STrengthening the Reporting of OBServational studies in Epidemiology)
Systematic Reviews and Meta-Analyses	<b>PRISMA</b> (the Preferred Reporting Items for Systematic Reviews and Meta-Analyses)
Systematic Reviews and Meta-Analyses Protocol	<b>PRISMA-P</b> (the Preferred Reporting Items for Systematic Reviews and Meta-Analyses - Protocol)
Experimental Animal Studies	<b>ARRIVE</b> (the Animal Research: Reporting of In Vivo Experiments)
Diagnostic Accuracy Studies	<b>STARD</b> (the Standards for Reporting Diagnostic Accuracy)
Qualitative Research	<b>SRQR</b> (the Standards for Reporting Qualitative Research)
	<b>COREQ</b> (COnsolidated criteria for Reporting Qualitative research: interviews and focus groups)
Methodological Studies (Translating and Adapting Tests)	<b>ITC</b> (International Test Commission) Guidelines for Translating and Adapting Tests ( <a href="https://www.intestcom.org/">https://www.intestcom.org/</a> )
Methodological Studies (Developing Tests)	<b>COSMIN</b> (COnsensus based Standarts fort he selection of Health Measurement Instruments- Study Design for Patient-reported outcome measurement instruments)
Case Report	<b>CARE</b> (Case Reporting)

---

\*Enhancing the QUALity and Transparency Of Health Research (equator network) (<https://www.equator-network.org/>)

\*\*The Lokman Hekim Health Sciences encourages the registration of all clinical trials (randomized and non-randomized) via ClinicalTrials.gov ([www.clinicaltrials.gov](http://www.clinicaltrials.gov)) or one of the registries of the WHO's International Clinical Trials Registry Platform (ICTRP: <http://www.who.int/ictcp/network/primary/en/index.html>). The name of the trial registry and the registration number together should be provided at the end of the abstract.

**Letter to the editor:** This type of manuscript discusses important observations, ignored aspects, or details lacking in a previously published article in Lokman Hekim Health Sciences. The article that is the subject of commentary must be properly cited within the manuscript. No abstract, keywords, tables, figures, images, or other media should be included. The text should be unstructured and is limited to 500 words. No more than 5 references will be accepted.

**Table 1.** Limitations for each manuscript type

Type of manuscript	Word limit	Abstract word limit	Reference limit	Table or figure limit
Editorial comments	1200	No abstract, keywords	10	No, tables, figures, images.
Original article	3000–3500	250 (Structured)	30	6
Brief report	1500	150	15	1 table and 3 figures
Narrative review	4000	250	40	6
Systematic review	4000	250 (Structured)	Depends on review frame (60 maximum)	6
Case report	1200	150 (Structured)	10	1 table, 2 figures or images
Letter to the editor	500	No abstract, keywords	5	No tables, figures, images
History of medicine	1500	250	depends on review frame (40 maximum)	6

### MANUSCRIPT SUBMISSION

The entire submission process for a manuscript is completed online through the self-explanatory online submission system through the following website: <https://lhhs.manuscriptmanager.net/>. An appropriate preparation and submission process will prevent delays in the review process of the manuscripts.

**Cover letter:** The cover letter should include the article name, article type, and the full name and address of the corresponding author. Corresponding author should declare the absence or presence of any conflict of interest in cover letter, and should affirm that the paper has not already been published, accepted, or is under simultaneous review for publication elsewhere in any language. For manuscripts that have been presented orally or as a poster, this must be stated on the title page with the date and the place of the presentation.

**Title page:** A title page should be submitted with all submissions and this page should include:

- The Turkish and English full title of the manuscript no more than 150 characters and a Turkish and English short title (running head) of no more than 50 characters
- Name, affiliation, ORCID ID number, e-mails and highest academic degree of the author(s)
- The statement of conflict of interest and funding information
- Name, address, phone number(s), and email address of the corresponding author
- Acknowledgment of the individuals who contributed to the preparation of the manuscript but who do not fulfill the authorship criteria

**Abstract:** A Turkish and English abstract is required with all submissions except editorial comments, images, and letters to the editor. Submitting a Turkish abstract is not compulsory for international authors.

**Keywords:** Each submission must be accompanied by a minimum of three and a maximum of six keywords for subject indexing included at the end of the abstract. The keywords should be selected from the National Library of Medicine, Medical Subject Headings database (<https://www.nlm.nih.gov/mesh/MBrowser.html>).

**Main document:** The main document should include the main text, the reference list, tables, and figure legends, respectively. Any information that may indicate an individual or institution should be excluded from the main document. Statistical analysis should be conducted in accordance with the guidelines on reporting statistical data in medical journals [Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *Br Med J* 1983; 7; 1489-93 and Lang T, Altman D. Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A (editors). *Science Editors' Handbook*, European Association of Science Editors, 2013.]. The software used for statistical analysis must be described. All references, tables, and figures should be referred to within the main text, and they should be numbered consecutively in the order they are referred to within the main text. Limitations and drawbacks of original articles should be mentioned in the Discussion section before the conclusion paragraph.

### Tables

Tables should be uploaded as separate file in word format and not embedded in the main text. They should be numbered consecutively in the order they are referred to within the main text. A descriptive title must be placed above the tables. Abbreviations used in the tables should be defined below the table with footnotes, even if they are defined within the main text.

### Figures and figure legends

Figures, graphics, and photographs should be submitted as separate files in TIFF or JPEG format through the article submission system. The files should not be embedded in a Word document or the main document. Thick and thin arrows, arrowheads, stars, asterisks, and similar marks

can be used on the images to support figure legend. Any information within the images that may identify an individual or institution should be blinded. The minimum resolution of each submitted figure should be 300 DPI. Figure legends should be listed at the end of the main document.

## References

The editorial team may further request that the authors cite relatively recently published articles (preferably within the last 5 years) in their manuscripts, with the exception of papers e.g., concentrating on the history of Medicine and Health Sciences. Reference listings must be in accordance with ICMJE standards and numbered consecutively at the very end of the manuscript in the order in which they are mentioned in the text. If an ahead of print publication is being cited the DOI number should be provided. Authors are responsible for the accuracy of the references. Journal titles should be abbreviated in accordance with the journal abbreviations in Index Medicus/ Medline/PubMed (for journal abbreviations please consult the List of Journals indexed for MEDLINE, published annually by NLM). provided there are 6 or less authors, all the authors should be enlisted. If there are 7 or more authors, the first 6 authors should be listed followed by "et al". In the main text of the manuscript, all the references should be cited using Arabic numbers in parentheses. Last, it is essential that all the references listed should be in English (subject titles and journal names should be in English). The subject titles in a language other than English and the journal names other than English need to be given in parentheses at the end. The reference styles for different types of publications are presented in the following examples in this regard:

**Journal article:** Wu CC, Econs MJ, DiMeglio LA, Insogna KL, Levine MA, Orchard PJ, et al. Diagnosis and management of osteopetrosis: Consensus guidelines from the Osteo-petrosis Working Group. *J Clin Endocrinol Metab* 2017;102(9):3111–23.

Armstrong N, Ryder S, Forbes C, Ross J, Quek RGW. A systematic review of the international prevalence of BRCA mutation in breast cancer. *Cecconi M, Evans L, Levy M, Rhodes A. Sepsis and septic shock. Lancet* 2018;392(10141):7587.

Aslan MH, Vural M. Gram Negative Bacteria Isolated from Blood Cultures and Their Antibiotic Susceptibility. *Trends Surg Sci* 2025;4(1):11-17. [In Turkish]

**Epub ahead-of-print article:** Huyut MA. Kidney Injury Molecule-1 Is Associated with Contrast-Induced Nephropathy in Elderly Patients with Non-STEMI. *Arq Bras Cardiol.* 2021 Mar 29;S0066-782X2021005005201. Portuguese, English. doi: 10.36660/abc.20200172. Epub ahead of print. PMID: 33787767.

**Manuscript published in electronic format:** T.C. Ministry of Health, General Directorate of Public Health. COVID-19 (SARS-CoV2 Infection) Guide (Science Board Study) March 25, 2020. [www.hsgm.saglik.gov.tr](http://www.hsgm.saglik.gov.tr)

**Book section:** Suh KN, Keystone JS. Malaria and babesiosis. Gorbach SL, Barlett JG, Blacklow NR, editors. *Infectious Diseases*. Philadelphia: Lippincott Williams; 2004.p.2290-308.

**Conference proceedings:** Bengissson S. Sothemin BG. Enforcement of data protection, privacy and security in medical informatics. In: Lun KC, Degoulet P, Piemme TE, Rienhoff O, editors. *MEDINFO 92. Proceedings of the 7<sup>th</sup> World Congress on Medical Informatics*; 1992 Sept 6-10; Geneva, Switzerland. Amsterdam: North-Holland; 1992. pp.1561-5.

**Scientific or technical report:** Cusick M, Chew EY, Hoogwerf B, Agrón E, Wu L, Lindley A, et al. Early Treatment Diabetic Retinopathy Study Research Group. Risk factors for renal replacement therapy in the Early Treatment Diabetic Retinopathy Study (ETDRS), Early Treatment Diabetic Retinopathy Study *Kidney Int*: 2004. Report No: 26.

## REVISIONS

When submitting a revised version of a paper (include a clean copy and a highlighted copy), the author must also upload the required files (Title page; Main clean text; Main highlighted text; Tables in Word format; Figures in TIFF or JPEG format) and must submit a detailed "Response to reviewers" that replies to each issue point by point raised by the reviewers and indicates where changes can be found (each reviewer's comment, followed by the author's reply and line number where changes have been made). Revised manuscripts must be submitted within 30 days from the date of the decision letter. If the revised version of the manuscript is not submitted within the allocated time, the revision option will be automatically withdrawn. If the submitting author(s) believe that additional time is required, they should request this extension within the initial 30-day period.

## PUBLICATION PROCESS

Accepted manuscripts are copy edited for grammar, punctuation, format, and clarity. Once the publication process of a manuscript is completed, it is published online on the journal's webpage as an ahead-of-print publication before it is included in the scheduled issue. A PDF proof of the manuscript is sent to the corresponding author and their publication approval is requested within 2 days of receipt of the proof.

Accepted manuscripts will be made available and citable online as rapidly as possible. The stages of publication are as follows;

**Uncorrected publication:** A PDF of the final, accepted (but unedited and uncorrected) paper will be published online on the journal web page under the "Accepted Articles" section. A DOI will be assigned to the article at this stage.

**Ahead-of-print publication:** After copy editing, typesetting, and review of the resulting proof, the final corrected version will be added online in the "Ahead-of-Print" section.

**Final publication:** The final corrected version will appear in an issue of the journal and added to the journal website. To ensure rapid publication, we ask authors to provide your publication approval during the proofreading process as quickly as possible, and return corrections within 48 hours of receiving the proof.

## SUBMISSION CHECKLIST (2026)

Use this checklist to prepare and finalize your manuscript for submission.

### Formatting requirements

- Font & Spacing: 12-point Times New Roman, 1.5 line spacing, 2.54 cm (1 inch) margins
- Headings: Main and subheadings in 12-point bold
- Spacing: Single space after sentences
- Style: Avoid bold for emphasis; use italics for foreign words/concepts
- Numbers: Write out one to ten unless a unit of measurement.
- Paragraphs: Single hard return; no tabs/indents.
- Technical Settings: No auto-hyphenation, headers, or footers
- Page & Line Numbers: Require
- Language: US English

### Required documents

#### 1. Cover letter

- o Article title & type, corresponding author's details
- o Conflict of interest statement
- o Ethics approval and/or patient consent
- o Funding information
- o Data availability statement

#### 2. Title page

- o Article type, title, running title
- o All authors' details and ORCID ID
- o Conflict of interest and funding statements
- o Corresponding author's contact details
- o Acknowledgments for non-author contributors
- o Conference presentation details (if applicable)
- o Financial or other supports
- o Word counts (abstract & main text)

### Main manuscript

- Title, abstract, and keywords
- Text with required subheadings per journal guidelines
- Correct figure/table citations in-text
- References formatted as per journal style
- Descriptive legends, figures, tables
- Figures and tables uploaded separately
- Required forms (Author Contribution, Copyright, ICMJE Disclosure)

### Final checks

- Statistical Analysis: Verified
- Language & Grammar: Use US English spell check.
- References: Cited in-text and listed correctly
- Permissions: Obtained for reproduced figures, tables, and photograph
- Ethical Approval & Informed Consent: Included for human subject research
- Abbreviations: Define
- Compliance: Journal policies followed

### REVISED MANUSCRIPT RE-SUBMISSION

1. Title page
2. Clean revised version of the main text (without line numbers).
3. Highlighted main text: A revised version of the manuscript with changes made in bold, underlined or highlighted (adding page and line numbers)
4. Tables in Word format should be separately uploaded
5. Figures in TIFF or JPEG format should be separately uploaded
6. Response to Reviewers: a point-to-point reply to the reviewers' comments that addresses each comment of each reviewer. Please be as specific as possible in your response to the reviewer(s).
7. If you have not already done so, please download the submission statements, fill them in and have each co-author sign the form(s).

## CONTENTS/ İÇİNDEKİLER

Special Issue: Artificial Intelligence in Medicine and Health Sciences

### EDITORIAL COMMENT

- The Artificial Intelligence Revolution in Medicine and Health Sciences – From Historical Roots to Future Therapeutics** ..... 177–178  
Şirin H

### ORIGINAL ARTICLES

- Thematic Analysis of ChatGPT’s Menopause-Related Patient Education Responses from a Women’s Health Nursing Perspective** ..... 179–187  
Bulut H, Alotaibi S
- Identification of Risk Factors for Type 2 Diabetes Mellitus: A Machine Learning Approach** ..... 188–195  
Budak S, Karacan Y, Bacak İ, Özer Ş
- The Use of Artificial Intelligence in Medical Education: A Comparative Analysis of Theoretical Exam Performance between ENT Residents and ChatGPT-4o** ..... 196–202  
Doğan Karataş T, Aksoy A, Bora A, Doğan M
- Evaluation of Chat Generative Pretrained Transformer-5 and the Cameriere Method in Dental Age Estimation** ..... 203–211  
Cömert H, Özdemir M, Ataç AS
- Use of Artificial Intelligence–Based Chatbots and Self-Efficacy Among Students Taking a Surgical Nursing Course** ..... 212–220  
Yeşilyaprak Karaca T, Özsoy H, Gök F
- Accuracy of Artificial Intelligence-Generated References in Dental Trauma Management** ..... 221–227  
Yıldırım Manav E, Özdemir M
- Bibliometric Trends of Artificial Intelligence Applications in Pediatric Dentistry: A Scopus and Web of Science Analysis (2000–2024)** ..... 228–236  
Çerçi Akçay H
- Determining the Artificial Intelligence Literacy and Attitudes toward Artificial Intelligence of Oncology Nurses: A Descriptive-correlational Study** ..... 237–245  
Karakuş Z, Semerci Şahin R, Babadağ Savaş B, Bağçivan G
- Assessment of Artificial Intelligence Chatbots’ Information Quality on Home Dental Bleaching: A Comparative Study** ..... 246–254  
Özişçi Ö, Gökmen El
- Evaluation of AI Chatbots in Tooth Avulsion Management According to the International Association of Dental Traumatology Guidelines** ..... 255–263  
Özdemir M, Yıldırım Manav E
- Publication Trends and Citation Patterns of Artificial Intelligence and Large Language Model Research in Orthopedics and Traumatology: A Türkiye-based Bibliometric Analysis** ..... 264–271  
Gezer MC, Demir M
- A Bibliometric Analysis of Artificial Intelligence and Simulation Technologies in Dental Education** ..... 272–281  
Demirbaş HG
- Evaluation of ChatGPT Responses to Questions About Joint Health and Physiotherapy in Patients with Hemophilia** ..... 282–289  
Tat NM, Tat AM

**SYSTEMATIC REVIEWS**

**Artificial Intelligence in Unexplained Infertility: A Systematic Review of Machine Learning-Based Predictive Models** ..... 290–302  
Tahta T, Aydoğan Aşır Ö

**Artificial Intelligence in Neurosurgical Education: A Systematic Review of Technical Skills Training, Clinical Reasoning, and Surgical Planning**..... 303–312  
Şahin ÖS, Dinç S

**Mapping Artificial Intelligence-based Assessment Domains in Pediatric Occupational Therapy: A Scoping Review** ..... 313–322  
Kolit Z, Alataş DM, Ayber BE

**A Systematic Review and Meta-Analysis of Diagnostic Performance Comparison between DeepSeek and Physicians** ..... 323–333  
Zeng J, Zhu X, Liu X, Shen S, Li S, Cao S

**Artificial Intelligence Applications in Pediatric Genetics: Clinical and Research Implications and Ethical Considerations**..... 334–340  
Öztürk Yılmaz Ş

**INVITED REVIEW**

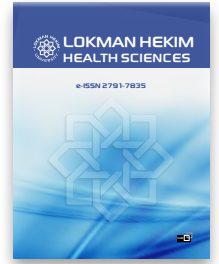
**Artificial Intelligence as a Managerial Decision Support Infrastructure in Hospitals: A Governance Framework** ..... 341–351  
Şahin İ, Zengul FD

**REVIEWS**

**Integrating Artificial Intelligence into Assisted Reproduction Technologies: Current Applications and Future Directions – A Narrative Review** ..... 352–359  
Ercenk R, Yıldırım Karaca S

**Artificial Intelligence–Driven Weight Management: Current Evidence and Clinical Implications** ..... 360–371  
Servi NN

**The Governance of Artificial Intelligence in Healthcare: Ethical Foundations, Legal Challenges, and Implementation Realities**..... 372–382  
Yılmaz M



# The Artificial Intelligence Revolution in Medicine and Health Sciences – From Historical Roots to Future Therapeutics

 Hülya Şirin

Department of Public Health, Gülhane Faculty of Medicine, University of Health Sciences, Ankara, Türkiye

This special issue brings together research at the intersection of artificial intelligence and the health sciences. Our focus spans artificial intelligence's (AI's) historical roots in medicine, its growing role in diagnosis and clinical practice, and the questions it raises for the future of patient care. Throughout, we aim to engage with these developments critically and responsibly – drawing on recent literature while keeping ethical considerations at the center of the discussion.

## A Historical Perspective: From Rule-Based Systems to Deep Learning

AI's place in medicine was not built overnight. The groundwork was laid decades ago, beginning in the 1970s with early expert systems such as MYCIN – a program developed to identify bacterial infections and suggest appropriate antibiotic treatments.<sup>[1]</sup> These rule-based tools were designed to support clinical decision-making, but their limited computing power and scarce data kept them largely within research settings. Even so, they introduced a foundational idea: that algorithms could play a meaningful role in medical reasoning. That idea has since grown considerably. With the rise of big data and modern computing, machine learning and deep learning have moved from the laboratory into everyday clinical practice.

## Transformation in Screening and Diagnostic Processes

Among AI's most consequential contributions to medicine are its abilities to process images and recognize complex patterns with remarkable consistency. In fields such as radiology, pathology, and emergency medicine, where data volumes are high and the margin for error is low, AI systems have demonstrated a reliable capacity to detect subtle findings that might otherwise be missed. A recent review from Türkiye reports that AI-based imaging tools reach accuracy rates of 85–90% in identifying conditions through X-rays and computed tomography scans, and show comparable performance in emergency triage.<sup>[2]</sup>

Yet the value of these tools extends beyond patients who are already symptomatic. Population-level screening – identifying individuals at risk before disease becomes clinically apparent – represents one of the most promising applications of predictive AI. By drawing on electronic health records and genetic data, these models can flag risk factors early, supporting the kind of preventive approach that has long been central to public health. Embedding such tools within national health systems could meaningfully strengthen early detection and reduce the long-term burden on healthcare infrastructure.

**Cite this article as:** Şirin H. The Artificial Intelligence Revolution in Medicine and Health Sciences – From Historical Roots to Future Therapeutics. Lokman Hekim Health Sci 2026;6(2):177–178.

**Correspondence:** Hülya Şirin, M.D. Sağlık Bilimleri Üniversitesi, Gülhane Tıp Fakültesi, Halk Sağlığı Anabilim Dalı, Ankara, Türkiye

**E-mail:** hulya.sirin@sbu.edu.tr **Submitted:** 05.05.2026 **Accepted:** 11.05.2026 **Available Online:** 02.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



## AI in Treatment: Personalized Medicine and Preventive Care

AI has also begun to reshape how treatment itself is delivered. The standardized, population-wide approach to care is increasingly being supplemented – and in some areas replaced – by personalized strategies tailored to the individual. Deep learning models can integrate a patient's biomarkers, genetic profile, and environmental context to support more targeted treatment decisions.<sup>[3]</sup> In parallel, AI-driven health platforms are being developed to support continuous monitoring, stress management, and proactive intervention, with particular attention to underserved areas such as elder care and social isolation.<sup>[4]</sup> In oncology and surgery, algorithms are contributing to immunotherapy response prediction and the optimization of robotic procedures, offering more precise and safer pathways for patients.

## Future Projections: Multimodal AI and Educational Readiness

Looking further ahead, AI is poised to accelerate medical research in ways that were until recently impractical. Drug discovery, historically a process of years and substantial investment, is being compressed through AI systems capable of modeling protein structures and running molecular simulations at speed.<sup>[5]</sup> The next step will likely be the widespread adoption of multimodal models – systems that can simultaneously interpret genetic data, laboratory results, imaging, and clinical history to produce a more complete picture of the patient.<sup>[6]</sup>

These advances, however, come with responsibilities that the field cannot afford to overlook. Questions of algorithmic bias, data governance, and clinical accountability require clear regulatory responses and sustained institutional attention. Equally important is the preparation of the next generation of healthcare professionals. Research suggests that medical and nursing students' engagement with AI is closely tied to their broader orientation toward innovation, underscoring the need to incorporate AI literacy into healthcare education as a matter of priority.<sup>[7]</sup>

The studies gathered in this special issue reflect this breadth – spanning diagnostic innovation, therapeutic application,

ethical inquiry, and educational readiness. We hope they contribute meaningfully to the ongoing conversation about how AI can be integrated into medicine in a way that is evidence-based, ethically grounded, and genuinely centered on patient care.

We wish you an enlightening and thought-provoking reading experience.

**Ethics Committee Approval:** Not applicable.

**Conflict of Interest:** None declared.

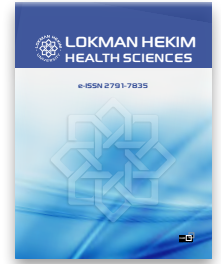
**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** Artificial intelligence (AI)-assisted technologies were used only for language editing.

**Peer-review:** Double blind peer-reviewed.

## References

1. Hirani R, Noruzi K, Khuram H, Hussaini AS, Aifuwa EI, Ely KE, et al. Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities. *Life (Basel)* 2024;14(5):557. [\[CrossRef\]](#)
2. Berikol GB, Kanbakan A, İlhan B, Doğanay F. Mapping artificial intelligence models in emergency medicine: A scoping review on artificial intelligence performance in emergency care and education. *Turk J Emerg Med* 2025;25(2):67-91. [\[CrossRef\]](#)
3. Fahim YA, Hasani IW, Kabba S, Ragab WM. Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives. *Eur J Med Res* 2025;30(1):848. [\[CrossRef\]](#)
4. Gocer H, Durukan AB, Gun T. Integrating artificial intelligence into personalized preventive medicine: addressing social isolation and elderly care. *JMA J* 2025;9(1):52. [\[CrossRef\]](#)
5. Mohajer-Bastami A, Moin S, Ahmad S, Ahmed AR, Pouwels S, Hajibandeh S, et al. Artificial intelligence in healthcare: applications, challenges, and future directions. A narrative review informed by international, multidisciplinary expertise. *Front Digit Health* 2025;7:1644041. [\[CrossRef\]](#)
6. Thomas KS, Edpuganti S, Puthooran DM, Thomas A, Joy A, Latheef S. Artificial intelligence in modern clinical practice (Review). *Med Int (Lond)* 2025;6(1):5. [\[CrossRef\]](#)
7. Kaya G, Büyükyılmaz F, Çulha Y, Akyürek P. Investigation of the relationship between medical artificial intelligence readiness and individual innovativeness levels in nursing students. *Nurse Educ Today* 2025;151:106721. [\[CrossRef\]](#)



# Thematic Analysis of ChatGPT's Menopause-Related Patient Education Responses from a Women's Health Nursing Perspective

Hatice Bulut<sup>1</sup>, Samar Alotaibi<sup>2</sup>

<sup>1</sup>Division of Obstetrics and Gynecologic Nursing, Department of Nursing, Süleyman Demirel University Faculty of Health Sciences, Isparta, Türkiye

<sup>2</sup>Department of Maternal and Child Health Nursing, College of Nursing, Shaqra University, Al Dawadmi, Saudi Arabia

## Abstract

**Introduction:** This study aimed to evaluate ChatGPT's patient education content on menopause from the perspective of women's health nursing and to identify emerging themes in its responses.

**Methods:** A qualitative design was employed. Fifteen scenarios covering common nursing counselling topics in menopause were developed and presented to ChatGPT with instructions to use "patient-friendly language." The responses were analysed using Braun and Clarke's six-phase thematic analysis method.

**Results:** Four main themes emerged: (1) Symptom-focused practical suggestions, (2) lifestyle modification and preventive health, (3) psychosocial support and relationships, and (4) safety and compliance. Most responses demonstrated empathy but lacked sufficient scientific references and evidence-based support.

**Discussion and Conclusion:** ChatGPT can provide useful and understandable suggestions for patient education in menopause from a nursing perspective. However, improvements are needed in providing evidence-based references and addressing certain topics (e.g., sexual health) more comprehensively. Women's health nurses should verify the accuracy of AI-generated content before using it in patient education. These findings highlight the importance of developing nurses' skills in critically appraising, validating, and appropriately integrating AI-generated educational content into women's health counselling and patient education.

**Keywords:** Artificial intelligence; ChatGPT; Menopause; Patient education; Thematic analysis; Women's health nursing

Menopause is a natural phenomenon defined as the absence of menstruation for one year due to the depletion of ovarian function. The onset of natural menopause usually occurs between the ages of 45 and 55 years worldwide. It is estimated that the number of

postmenopausal women worldwide will reach 1.2 billion by 2030.<sup>[1]</sup> Symptoms of menopause vary and arise from the complex interplay of biological, psychological, and social factors. A decline in oestrogen levels may result in a wide range of symptoms, including hot flashes, night

**Cite this article as:** Bulut H, Alotaibi S. Thematic Analysis of ChatGPT's Menopause-Related Patient Education Responses from a Women's Health Nursing Perspective. *Lokman Hekim Health Sci* 2026;6(2):179–187.

**Correspondence:** Hatice Bulut, PhD. Süleyman Demirel Üniversitesi, Sağlık Bilimleri Fakültesi, Doğum ve Kadın Hastalıkları Hemşireliği Anabilim Dalı, Isparta, Türkiye

**E-mail:** haticebulut@sdu.edu.tr / hatice.phd.sheffield@gmail.com

**Submitted:** 29.09.2025 **Revised:** 07.01.2026 **Accepted:** 27.01.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



sweats, fatigue, sleep disturbances, depression, cognitive difficulties, reduced sexual desire, and joint and muscle pain. Vasomotor symptoms, such as hot flashes and night sweats, are the most frequently reported, affecting up to 80% of women.<sup>[2]</sup> Although menopause is a shared biological process, sociocultural contexts influence women's experiences, resulting in varied approaches to coping with the transition.<sup>[3]</sup> The way women experience menopause can be negatively shaped by their health status, life history, and sociocultural environment.<sup>[4]</sup>

Nurses play a key role in providing evidence-based guidance, counselling, and support to women as they navigate the menopause transition.<sup>[5]</sup> An advanced nurse practitioner-led menopause clinic significantly improved overall quality of life and alleviated nearly all assessed menopausal symptoms, suggesting its value in expanding timely menopause care within the workforce.<sup>[6]</sup> Addressing menopausal symptoms and improving women's self-perception of aging require a holistic approach, in which clinicians and nurses integrate education, psychological support, and personalized care. By providing accurate, evidence-based information about the physiological changes associated with aging and menopause, healthcare professionals can help dispel stigma and prevent misconceptions.<sup>[7]</sup>

Recently, there has been a shift in healthcare toward the use of technology and artificial intelligence (AI). AI is increasingly seen as a valuable aid for physicians dealing with the complexities of menopause, presenting both opportunities and challenges.<sup>[8,9]</sup> Natural language processing offers a promising approach for providing customized recommendations for symptom management, menopausal hormone therapy, and alternative therapies, thereby optimizing care according to individual medical history and preferences.<sup>[10]</sup> However, a recent study found that, compared with an artificial intelligence-based program, a health-promotion-oriented intervention was more effective in improving menopausal women's knowledge and quality of life.<sup>[11]</sup> Therefore, achieving a balance between the use of AI and the preservation of human knowledge and empathy is critical.<sup>[12]</sup>

In light of this, it is important to evaluate the content of health education on menopause to ensure it supports informed decision-making, promotes patient-centered care, and addresses the complex biological, psychological, and social dimensions of the menopausal transition. Although several studies have explored the use of AI and chatbots in healthcare and patient education, few

have examined the quality of AI-generated educational content from a nursing perspective. To our knowledge, no studies have evaluated ChatGPT's menopause-related information through a women's health nursing lens. This study addresses this gap by analyzing ChatGPT's menopause education responses using scenario-based thematic analysis to explore its potential contributions and limitations in women's health counselling. Therefore, this study aimed to evaluate ChatGPT's patient education content on menopause from the perspective of women's health nursing and to identify emerging themes in the responses.

## Materials and Methods

### Study Design

This study was a qualitative study conducted digitally, independent of the clinical setting. The study design was informed by both authors' long-standing expertise in menopause and doctoral-level academic experience in women's health nursing and menopause. The study aimed to contextually evaluate ChatGPT-5 responses to patient education and counselling scenarios related to menopause. The scenarios were developed based on common menopausal symptoms and care recommendations outlined by the World Health Organization,<sup>[13]</sup> the National Health Service,<sup>[14]</sup> and the North American Menopause Society.<sup>[15,16]</sup>

### Type of Research

The research was based on a qualitative thematic analysis approach. For this purpose, 15 unique patient-language scenarios were developed, drawing on the menopause literature and a nursing care perspective. The scenarios were structured to cover common symptoms and care areas during menopause (e.g., vasomotor symptoms, sleep problems, vaginal dryness, osteoporosis risk, and psychosocial impacts).<sup>[13-16]</sup>

### Data Collection

Fifteen scenarios prepared by the researcher were individually submitted to ChatGPT in Turkish, and the model was specifically asked to use "patient-friendly language" in its responses. Only one response was generated for each scenario, and the resulting texts were transcribed into a written document without any modifications. The responses were systematically recorded and later incorporated into the qualitative analysis process. A full list of scenarios is presented in Table 1 and Appendix 1.

**Table 1.** Menopause scenarios included in the study

Scenario no	Scenario topic
1	Hot flush management (non-medication)
2	Night sweats
3	Vaginal dryness and pain
4	Urinary incontinence
5	Decreased libido
6	Osteoporosis prevention
7	Fall risk reduction
8	Heart health
9	Weight management
10	Sleep problems
11	Mood swings
12	Counselling about HRT
13	Herbal supplements
14	Workplace menopause support
15	Family support

HRT: Hormone replacement therapy.

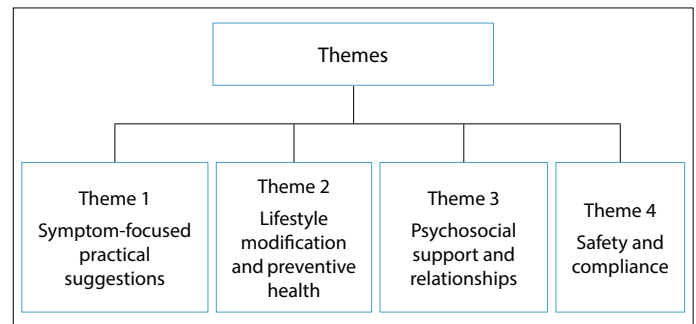
## Data Analysis

The responses obtained in the study were analysed using Braun and Clarke's thematic analysis method (2006).<sup>[17]</sup> In the first stage, all data generated by ChatGPT-5 were read multiple times by the researcher to ensure a comprehensive familiarity with the content. In the second stage, meaningful sections of the responses were systematically coded, and these codes were grouped according to shared content. Potential themes were then developed from the codes and subsequently reviewed and reorganized to ensure content consistency. In the final stage, the themes were finalized and interpreted within the context of a nursing care perspective. A sample coding table is presented in Table 2.

All scenarios were submitted in Turkish, and ChatGPT outputs were generated in Turkish. The Turkish outputs were retained verbatim as the source material. To facilitate collaborative analysis with an English-speaking co-researcher, the outputs were translated into English by the research team. The translations were checked against the original Turkish text during coding and theme development to ensure fidelity of meaning, and any discrepancies were resolved by consensus.

## Rigour and Reflexivity

Scientific rigour and reliability were ensured in this study through various strategies. Both the first and second researchers have worked in the field of menopause for more

**Figure 1.** Thematic map of the four main themes identified in the study.

than 10 years and completed their doctoral dissertations in this area. In addition, given the researcher's academic work on qualitative research methods and menopause, a reflexive approach was adopted, recognizing that her personal knowledge and experiences could influence the research process.

Specifically, during the data analysis process, systematic self-reflection was undertaken to reduce the risk of bias, in accordance with Braun and Clarke's (2006) thematic analysis approach.<sup>[17]</sup> All steps were implemented transparently. A reflexive approach, supported by both the researcher's field expertise and feedback from the second researcher, was maintained throughout the research process, thereby maximizing the reliability, validity, and scientific contribution of the findings.

## Ethical Approval

This research did not require ethics committee approval because it did not involve human participants. The study was conducted solely using responses from an AI model and did not involve any personal data.

## Results

The data obtained in this study revealed that recommendations addressing women's experiences during menopause are multidimensional. The thematic analysis identified four main themes, which were then organized into subcategories. Overall, the responses consisted of concrete suggestions directly related to each scenario and included practical tips that menopausal women could easily implement in their daily lives (Fig. 1).

### Theme 1 – Symptom-Focused Practical Suggestions

A substantial proportion of the responses offered practical solutions and suggestions for alleviating menopausal symptoms in daily life. These suggestions were symptom-specific, cost-effective, and readily applicable.

**Table 2.** Sample coding table

Scenario no	Quote from text	Code	Theme	Description
1	"Keep your bedroom cool (18–20°C may be ideal)."	Symptom management	Theme 1 – Symptom-Focused Practical Suggestions	Direct suggestion for alleviating vasomotor symptoms.
6	"Increase your daily calcium intake: milk, yogurt, cheese, almonds, green leafy vegetables."	Lifestyle suggestion	Theme 2 – Lifestyle Modification and Preventive Health	Nutritional suggestion against the risk of osteoporosis.
5	"Communicate openly and honestly with your partner; share your feelings and expectations."	Psychosocial support	Theme 3 – Psychosocial Support and Relationships	Communication-based suggestion for decreased sexual desire.
12	"Like any medication, it has risks: the risk of breast cancer, blood clots, and stroke may increase."	Risk warning	Theme 4 – Safety and Compliance	Warning that draws attention to the potential risks of HRT.

HRT: Hormone replacement therapy.

### Subtheme 1.1: Management of Physical Symptoms

In scenarios related to vasomotor symptoms (hot flushes and night sweats), the responses suggested practical strategies such as regulating ambient temperature, choosing lightweight clothing, and taking cool showers or drinking cold beverages. The responses indicated that hot flushes disrupt daily quality of life, making cooling measures during sleep particularly important.

*"For hot flushes, keep a bottle of cold water by your bed at night, and have a thin spare pyjama set ready if needed." (Scenario 1)*

For vasomotor symptoms (hot flushes and night sweats), the responses also offered low-cost and practical recommendations, such as maintaining a cool environment, wearing cotton clothing, drinking cold beverages, and using soothing measures during sleep.

*"Try to keep your environment cool during the day and while sleeping." (Scenario 1)*

*"Keep your bedroom cool (18–20°C may be ideal)." (Scenario 2)*

The responses highlighted the use of lubricants, hygiene recommendations, and pelvic floor exercises for vaginal dryness, while Kegel exercises, fluid management, and weight control were emphasized for urinary incontinence.

*"Use water-based lubricants; be cautious, as oil-based products can weaken condoms." (Scenario 3)*

*"Do Kegel exercises three times a day, 10 repetitions each: tighten the muscles as if trying to stop urine flow, hold for 5 seconds, then release." (Scenario 4)*

### Subtheme 1.2: Behavioral Regulation and Psychosocial Approaches

Responses to the scenarios frequently emphasized the need to avoid foods and beverages thought to exacerbate menopausal symptoms. Spices, caffeine, alcohol, and hot drinks, in particular, were repeatedly cited as triggers of hot flushes.

*"Avoid spicy foods, hot beverages, caffeine, and alcohol, as they can trigger hot flushes." (Scenario 1)*

The responses also emphasized that stress can exacerbate symptoms and that methods such as relaxation exercises, deep breathing techniques, and yoga can be beneficial. This suggests that symptoms are linked not only to biological but also to psychological processes.

*"Find ways to reduce your stress: deep breathing exercises, gentle yoga, meditation." (Scenario 1)*

*"Relaxation activities such as yoga, meditation, and breathing exercises may positively influence sexual desire." (Scenario 5)*

*"Remember that stress, fatigue, and emotional state play an important role in sexual desire; make time for rest." (Scenario 5)*

### Theme 2 – Lifestyle Modification and Preventive Health

This theme encompasses lifestyle recommendations aimed at reducing long-term health risks and improving quality of life during menopause. Responses are grouped around both health-promoting approaches based on nutrition and physical activity and preventive measures such as avoiding harmful habits and maintaining good sleep hygiene.

### **Subtheme 2.1: Healthy Nutrition and Regular Physical Activity**

The responses recommended that women consume foods rich in calcium and vitamin D, increase their intake of fruits and vegetables, and choose whole grains and healthy fats. These recommendations appear to reflect a preventive approach to bone health and metabolic balance.

*"Increase your daily calcium intake: milk, yogurt, cheese, almonds, leafy green vegetables." (Scenario 6)*

Regular exercise was highlighted as beneficial for both cardiovascular and musculoskeletal health. Recommendations generally focused on activities that are easily accessible, such as walking, yoga, and light jogging.

*"Engage in at least 150 minutes of moderate-intensity walking or exercise per week." (Scenario 8)*

### **Subtheme 2.2: Avoiding Harmful Habits and Sleep Hygiene**

A substantial proportion of the responses identified smoking and alcohol use as factors that worsen menopausal symptoms and increase long-term health risks. Some responses made direct connections, such as smoking triggering cough and urinary incontinence.

*"If you smoke, quit; coughing can increase urinary incontinence." (Scenario 4)*

For women experiencing sleep problems, recommendations emphasized establishing a bedtime routine, reducing screen use, maintaining good sleep hygiene, and practicing relaxation techniques.

*"Stop using electronic devices at least one hour before sleep." (Scenario 10)*

### **Theme 3 – Psychosocial Support and Relationships**

The responses indicated that menopause is not only a biological process but also a psychosocial one. The suggestions generated by ChatGPT offered supportive strategies across a wide range of topics, from communication with partners and family to workplace conditions.

#### **Subtheme 3.1: Support in Intimate Relationships (Partner, Family, and Psychological Resilience)**

The impact of menopause on sexual desire and family dynamics was highlighted. In scenarios such as decreased sexual desire, the responses emphasized the importance of open communication and emotional sharing. It was noted that empathetic communication with a partner can improve relationship harmony.

*"Communicate openly and honestly with your partner; share your feelings and expectations." (Scenario 5)*

The responses also suggested that family members can be understanding of women experiencing menopause and provide support with household responsibilities. These suggestions indicate that menopause also affects family relationships.

*"Do not belittle or dismiss her feelings." (Scenario 15)*

Strategies such as journaling, exercise, breathing techniques, and, when necessary, professional psychological support were recommended for managing stress, anger, and emotional fluctuations.

*"Express your feelings by keeping a diary." (Scenario 11)*

#### **Subtheme 3.2: Support Mechanisms in Work and Social Life**

The effects of menopause on working life were highlighted, and the responses suggested that adjustments should be made to the work environment. Recommendations included improving ventilation, allowing cooling-off breaks, and adopting a flexible dress code. This underscores the need to address menopause as a process integrated with women's professional identities.

*"Ensure good ventilation in the work environment." (Scenario 14)*

### **Theme 4 – Safety and Compliance**

This theme encompasses both clinical-level medical risks and safety warnings, as well as recommendations appropriate to the cultural and environmental context of daily life.

#### **Subtheme 4.1: Clinical and Daily Safety**

The responses emphasized both the benefits of hormone replacement therapy (HRT) and its potential risks (e.g., breast cancer, thrombosis, and stroke). Similarly, it was noted that herbal products should be used with caution due to possible side effects and drug interactions.

*"Like any medication, it carries risks: it can increase the risk of breast cancer, blood clots, and stroke." (Scenario 12)*

*"If you are taking blood thinners, blood pressure medication, or hormone-related treatments, always consult your doctor first." (Scenario 13)*

Precautions that can be taken in the home environment (e.g., non-slip carpeting, good lighting, and well-organized walking areas) were highlighted, particularly in relation to the increased risk of falls with age.

"Organize walking areas at home; remove cables or rug edges that could cause tripping." (Scenario 7)

#### **Subtheme 4.2: Cultural and Local Adaptation**

The responses highlighted suggestions that were appropriate to the climate and living conditions in Turkish society. In particular, recommendations related to clothing choices (e.g., cotton underwear and comfortable clothing) and home arrangements (e.g., carpets and lighting) reflected adaptation to the local context.

*"Prefer cotton underwear; avoid tight and synthetic clothing." (Scenario 3)*

*"Take precautions at home to reduce the risk of falls (non-slip rugs, proper lighting)." (Scenario 6)*

### **Discussion**

This study examined ChatGPT's responses to scenarios related to menopause patient education from a women's health nursing perspective through thematic analysis. The findings indicate that an AI-based language model can provide empathetic, understandable, and practical advice for patient education related to menopause, but that it lacks context regarding sexual health, emotional support, and references to scientific sources. These findings are important because they provide an early assessment of how large-scale language models such as ChatGPT can serve as helpful tools in nursing practice and women's health education.

The four main themes arising from the thematic analysis demonstrate that the menopause experience is not merely a biological process but is also closely linked to psychological, social, and environmental factors. Our findings support the biopsychosocial model, which is widely accepted in the international literature on menopause management.<sup>[18,19]</sup> One of the most striking findings of our study is that ChatGPT addressed not only the biological but also the psychosocial dimensions of menopause, aligning with the biopsychosocial approach that underpins nursing. The ability of an AI system to address a health issue within its relational and social context suggests that this technology has the potential to go beyond superficial information transfer in patient education.

In this study, ChatGPT's practical recommendations for daily living, such as staying in a cool environment, choosing cotton and sweat-absorbing clothing, and avoiding triggers (alcohol, spices, and hot drinks), align with the literature indicating their effectiveness in managing mild hot flushes. Evidence also suggests that lifestyle changes can reduce

symptoms and that such environmental measures may be effective.<sup>[20,21]</sup> Furthermore, hot flushes have been shown to be physiologically triggered by small increases in body temperature resulting from a narrowing of the thermoneutral zone. This indicates that recommendations for regulating environmental temperature are scientifically grounded.<sup>[21]</sup>

The study found that psychological approaches, such as stress management, relaxation exercises, and breathing techniques recommended by ChatGPT, may play a supportive role in alleviating vasomotor symptoms. This finding is consistent with the results of a systematic review suggesting that psychological interventions, particularly cognitive behavioral therapy and relaxation techniques, may be effective in treating hot flushes in healthy postmenopausal women.<sup>[22]</sup>

The findings indicate that ChatGPT emphasizes healthy eating, exercise, and the avoidance of harmful habits in menopause management. The literature also emphasizes the importance of regular physical activity and a balanced diet rich in calcium and vitamin D in reducing the risk of postmenopausal osteoporosis and cardiovascular disease.<sup>[23,24]</sup> In this respect, ChatGPT's responses can be said to reflect a preventive health approach. However, the recommendations were mostly general and did not include details such as personalized exercise or nutrition plans.

A significant shortcoming in ChatGPT's responses is the lack of evidence-based source citations. Research suggests that the model can provide empathetic and understandable answers to patient questions, but the content is not always sufficient in terms of scientific accuracy.<sup>[25]</sup> Similarly, some systematic reviews have shown that although ChatGPT demonstrates high accuracy in certain medical fields, it can be inconsistent in specialized topics.<sup>[26]</sup> Therefore, it is important for nurses not to rely on AI-generated content directly, but to evaluate it by comparing it with existing guidelines.

The emphasis placed on partner, family, and workplace support in ChatGPT's responses is striking. This demonstrates that AI approaches menopause not only as a biological process, as emphasized in the literature, but also as a holistic biopsychosocial and cultural process that can affect women's sexual and intimate relationships.<sup>[27-29]</sup> Research has shown that menopausal women's experiences are shaped by deep and complex psychosocial dynamics, such as feeling a "duty" or "obligation" to meet their partner's needs in order to maintain the relationship, even when they themselves lack sexual desire.<sup>[27,28]</sup> However, ChatGPT's recommendations on this topic remain superficial.

Nevertheless, the safe use of AI in healthcare places significant ethical responsibilities on nurses. AI can provide generalized, non-personalized answers that may not apply to everyone because of biases in the data on which it is trained. At this point, the nurse's role is to critically assess the information provided by AI in order to prevent potential harm. In other words, nurses should personalize care by filtering AI-generated information and adapting it to the patient's specific situation.<sup>[30,31]</sup>

### Strengths, Limitations and Implications

One of the key strengths of this study is its evaluation of ChatGPT-5. While most previous studies have been limited to versions based on GPT-3.5 or GPT-4, this research demonstrates the current capabilities of AI-based chatbots. Accordingly, the findings not only provide a comparative perspective on earlier versions but also illustrate how current technology may contribute to patient education practices in women's health nursing. Furthermore, to our knowledge, this is the first study to examine ChatGPT's responses for menopause-related patient education from a women's health nursing perspective. In this respect, the study makes an original and pioneering contribution to the literature.

Another key strength of this study is the systematic, scenario-based evaluation of an AI-based tool such as ChatGPT in the context of menopause. The research examined responses not only across individual symptoms but also across multidimensional domains such as lifestyle, psychosocial relationships, and cultural context. This approach offers a unique contribution by demonstrating the scope of AI applications in women's health. Furthermore, the methodological framework, based on Braun and Clarke's (2006) six-stage thematic analysis, ensured that the findings were generated in a transparent and systematic manner.

The fact that the two researchers who conducted this study graduated from international universities, wrote their doctoral theses on menopause, and have published work on menopause can also be considered a strength of the study in terms of the perspective and analytical depth it provides.

The limitations of this study include the analysis of responses only in Turkish and the evaluation of responses from a single model. The exclusive use of Turkish-language scenarios may have influenced the nature and tone of ChatGPT's outputs, as linguistic and cultural nuances can shape how the model interprets and generates health-related information. In addition, because large language models are trained on extensive datasets that may contain linguistic, cultural, or contextual biases, the generated

responses may also reflect these limitations. Furthermore, this study was limited to content analysis; the actual effects of AI-generated responses on patients were not measured.

Future research could explore the effectiveness of ChatGPT-based educational interventions through experimental designs such as randomized controlled trials or user satisfaction surveys involving nursing students and menopausal women. These studies could assess the impact of AI-generated content on participants' knowledge, attitudes, and confidence in managing menopausal symptoms. In addition, future AI tools could be enhanced through integration with evidence-based guideline databases and digital health repositories (e.g., WHO and NICE), which would improve both the reliability and traceability of the information provided to patients.

In clinical practice, nurses can play a key role in ensuring that AI-generated educational materials are accurate and appropriate. To critically appraise such content, nurses should cross-check AI outputs against evidence-based clinical guidelines and national health recommendations, verify the reliability of cited information, and consider patients' cultural background and health literacy levels before using them. Furthermore, AI-based materials should be supplemented with nurse-led counselling and educational resources to ensure contextual relevance and accuracy. Strengthening nurses' digital health literacy and critical appraisal skills will be essential for the safe integration of AI tools into women's health education and counselling.

### Conclusion

In conclusion, ChatGPT has the potential to provide empathetic, actionable, and partially culturally appropriate recommendations for menopausal women. These findings suggest that ChatGPT can offer useful and understandable support for menopause patient education. From a women's health nursing perspective, ChatGPT may serve as a supportive tool, but its content requires careful verification for accuracy. Therefore, it is recommended that AI literacy courses be incorporated into nursing curricula and that healthcare professionals be trained to use this technology critically.

**Ethics Committee Approval:** The study did not involve human participants. Data was generated by artificial intelligence. Ethics committee approval was not required.

**Conflict of Interest:** The authors declare no conflicts of interest.

**Funding Statement:** The authors declared that this study has received no financial support.

**Acknowledgements:** The authors extend their appreciation to the OpenAI team.

**Use of AI for Writing Assistance:** ChatGPT was used for data collection. ChatGPT generated responses to pre-defined menopause scenarios (e.g., hot flush management, vaginal dryness, osteoporosis prevention). These AI outputs were thematically analysed by the research team and used as the primary qualitative data source for the study (as described in detail in the methods section).

**Data Availability Statement:** A data set will be available upon request.

**Author Contributions:** Concept: HB, SA; Design: HB, SA; Supervision: HB, SA; Materials: HB, SA; Data Collection or Processing: HB, SA; Analysis or Interpretation: HB, SA; Literature Search: HB, SA; Writing – original draft: HB, SA; Critical Reviews: HB, SA.

**Peer-review:** Double blind peer-reviewed.

## References

1. Afshari F, Bahri N, Sajjadi M, Mansoorian MR, Tohidinik HR. Menopause uncertainty: the impact of two educational interventions among women during menopausal transition and beyond. *Prz Menopauzalny* 2020;19(1):18-24. [Crossref]
2. El Khoudary SR, Greendale G, Crawford SL, Avis NE, Brooks MM, Thurston RC, et al. The menopause transition and women's health at midlife: a progress report from the Study of Women's Health Across the Nation (SWAN). *Menopause* 2019;26(10):1213-27. [Crossref]
3. Bahri N, Latifnejad Roudsari R. Moving from uncertainty toward acceptance: a grounded theory study on exploring Iranian women's experiences of encountering menopause. *J Psychosom Obstet Gynecol* 2020;41(2):154-64. [Crossref]
4. Talaulikar V. Menopause transition: Physiology and symptoms. *Best Pract Res Clin Obstet Gynaecol* 2022;81:3-7. [Crossref]
5. Flanagan MR, Fantasia HC. Comprehensive management of menopausal symptoms. *Nurs Womens Health* 2024;28(5):381-92. [Crossref]
6. Keye C. The Effect of an advanced nurse practitioner led menopause clinic on quality of life and menopausal symptoms. *Womens Health (Lond)* 2025;21:17455057251324573. [Crossref]
7. Wu J, Lu L, Jing L. Perspectives on improving self-perception of ageing and managing menopausal symptoms. *J Adv Nurs* 2026;82(3):2502-3. [Crossref]
8. Roa Diaz ZM, Muka T, Franco OH. Personalized solutions for menopause through artificial intelligence: Are we there yet? *Maturitas* 2019;129:85-6. [Crossref]
9. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380(14):1347-58. [Crossref]
10. Garg R, Munshi A. Revolutionizing menopause management: Harnessing the potential of artificial intelligence. *J Midlife Health* 2024;15(2):53-4. [Crossref]
11. Abd Elkader Mohamed W, Gharib Ahmed S, Abo Bakr Abd Ellatif M. Effectiveness of health promoting based program versus artificial intelligence based program on quality of life among menopausal women. *Egypt J Heal Care* 2024;15(2):1411-32. [Crossref]
12. Vargas-Hernandez VM. Artificial intelligence in menopause management. *Mathews J Gynecol Obstet* 2025;9(1):1-4. [Crossref]
13. World Health Organization (WHO). Menopause. World Health Organization. 2024. Available at: <https://www.who.int/news-room/fact-sheets/detail/menopause> Accessed Feb 2, 2025.
14. National Health Service (NHS). Menopause symptoms. NHS. Available at: <https://www.nhs.uk/conditions/menopause/symptoms/> Accessed Feb 10, 2025.
15. North American Menopause Society (NAMS). Menopause topics. The North American Menopause Society. Available at: <https://menopause.org/patient-education/menopause-topics> Accessed Feb 10, 2025.
16. Shifren JL, Gass ML; NAMS Recommendations for Clinical Care of Midlife Women Working Group. The North American Menopause Society recommendations for clinical care of midlife women. *Menopause* 2014;21(10):1038-62. [Crossref]
17. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006;3(2):77-101. [Crossref]
18. Melby MK, Lock M, Kaufert P. Culture and symptom reporting at menopause. *Hum Reprod Update* 2005;11(5):495-512. [Crossref]
19. Melby MK, Lampl M. Menopause, a biocultural perspective. *Annu Rev Anthropol* 2011;40:53-70. [Crossref]
20. Bansal R, Aggarwal N. Menopausal hot flashes: a concise review. *J Midlife Health* 2019;10(1):6-13. [Crossref]
21. Freedman RR. Physiology of hot flashes. *Am J Hum Biol* 2001;13(4):453-64. [Crossref]
22. Samami E, Shahhosseini Z, Elyasi F. The effects of psychological interventions on menopausal hot flashes: A systematic review. *Int J Reprod Biomed* 2022;20(4):255-72. [Crossref]
23. El Khoudary SR, Aggarwal B, Beckie TM, Hodis HN, Johnson AE, Langer RD, et al. Menopause transition and cardiovascular disease risk: implications for timing of early prevention: a scientific statement from the American Heart Association. *Circulation* 2020;142(25):e506-e32. [Crossref]
24. National Health Service (NHS). Things you can do - Menopause. NHS. Available at: <https://www.nhs.uk/conditions/menopause/things-you-can-do/> Accessed Feb 20, 2025.
25. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183(6):589-96. [Crossref]
26. Beheshti M, Toubal IE, Alaboud K, Almalaysha M, Ogundele OB, Turabieh H, et al. Evaluating the reliability of ChatGPT for health-related questions: a systematic review. *Informatics* 2025;12(1):1-23. [Crossref]
27. Bulut H. An interpretative phenomenological analysis of post-menopausal women's experiences of the menopause, its impact on their relationship and intimacy, and their coping methods [Doctoral dissertation]. United Kingdom: University of Sheffield; 2021.
28. Bulut H, Hinchliff S, Ali P, Piercy H. Women's experiences of intimate and sexual relationships during menopause: a qualitative synthesis. *J Clin Nurs* 2025;34(5):1543-54. [Crossref]

- 
29. Hunter M, Rendall M. Bio-psycho-socio-cultural perspectives on menopause. *Best Pract Res Clin Obstet Gynaecol* 2007;21(2):261-74. [\[Crossref\]](#)
30. Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform* 2019;28(1):128-34. [\[Crossref\]](#)
31. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. *Front Digit Health* 2021;3:645232. [\[Crossref\]](#)

## **Appendix 1. Scenarios for Menopausal Women**

---

### **Scenario 1 – Managing Hot Flashes (Non-Pharmacological)**

---

#### **Question:**

As a nurse, what non-medication advice would you give to a woman frequently experiencing hot flashes during menopause? Use patient-friendly language.

#### **Answer:**

- Try to keep your environment cool during the day and while sleeping.
  - Choose loose, light-coloured cotton clothing; avoid synthetic fabrics.
  - Avoid spicy foods, hot beverages, caffeine, and alcohol as they can trigger hot flashes.
  - Find ways to reduce stress: deep breathing exercises, gentle yoga, or meditation.
  - Take a 30-minute brisk walk daily; regular exercise helps reduce hot flashes and improves mood.
  - Keep a bottle of cold water by your bed at night, and have a thin spare pyjama set ready if needed.
  - Drink plenty of water to help maintain your body's fluid balance.
- 

### **Scenario 2 – Night Sweats**

---

#### **Question:**

As a nurse, what suggestions would you give to improve sleep quality for a woman experiencing night sweats during menopause?

#### **Answer:**

- Keep your bedroom cool (18–20°C may be ideal).
  - Use breathable cotton sheets and pyjamas.
  - Avoid heavy meals, caffeine, and alcohol before bedtime.
  - Take a warm shower before sleep to regulate your body temperature.
  - Keep a spare pyjama set and a thin towel next to your bed.
  - Establish a regular sleep routine; try to go to bed and wake up at the same time every day.
  - Practice relaxation techniques before sleep (e.g., the 4-7-8 breathing technique).
- 

### **Scenario 3 – Vaginal Dryness and Pain**

---

#### **Question:**

As a nurse, what non-medication recommendations would you offer to a woman experiencing vaginal dryness and pain during intercourse in the menopausal period?

#### **Answer:**

- You can use water-based lubricants; be cautious, as oil-based products can weaken condoms.
  - Regular sexual activity or pelvic floor exercises can improve blood circulation in the area and help reduce dryness.
  - Avoid using soaps or scented products in the genital area, as they can disrupt the pH balance.
  - Drink plenty of water to help maintain overall hydration.
  - Prefer cotton underwear; avoid tight or synthetic clothing.
  - If the dryness becomes very uncomfortable, consult your doctor about low-dose vaginal moisturizers.
- 

### **Scenario 4 – Urinary Incontinence**

---

#### **Question:**

As a nurse, what pelvic floor exercises and lifestyle changes would you recommend to a woman experiencing urinary incontinence during menopause?

#### **Answer:**

- Do Kegel exercises three times a day, 10 repetitions each: tighten the muscles as if trying to stop urine flow, hold for 5 seconds, then release.
  - Avoid doing this exercise during urination regularly; it's okay initially to identify the muscles, but don't make it a habit.
  - Drink plenty of fluids, but limit caffeinated and carbonated drinks as they may irritate the bladder.
  - If you're overweight, losing weight in a healthy way can help reduce symptoms.
  - Quit smoking if you smoke; chronic coughing can worsen incontinence.
  - Try not to hold your urine; go to the toilet at regular intervals.
- 

### **Scenario 5 – Low Sexual Desire**

---

#### **Question:**

As a nurse, what recommendations would you give to a woman experiencing decreased sexual desire after menopause, including psychosocial support?

#### **Answer:**

- Communicate openly and honestly with your partner; share your feelings and expectations.

- Remember that stress, fatigue, and emotional state play an important role in sexual desire; make time for rest.
- Physical intimacy that you both enjoy (hugging, massage) can help increase closeness.
- Use lubricants if vaginal dryness is an issue.
- Relaxation activities such as yoga, meditation, and breathing exercises may positively influence sexual desire.
- Take time for self-care to boost self-confidence.

---

### Scenario 6 – Preventing Osteoporosis

---

**Question:**

As a nurse, what nutrition, exercise, and lifestyle recommendations would you give to reduce the risk of osteoporosis after menopause?

**Answer:**

- Increase your daily calcium intake: milk, yogurt, cheese, almonds, leafy green vegetables.
- Have your vitamin D levels checked; make sure to get regular sunlight.
- Do weight-bearing exercises that support bone health, such as walking, light jogging, or climbing stairs.
- Avoid smoking and excessive alcohol consumption.
- Take precautions at home to reduce the risk of falls (non-slip rugs, proper lighting).

---

### Scenario 7 – Reducing Fall Risk

---

**Question:**

As a nurse, what home safety and exercise recommendations would you give to reduce the risk of falls after menopause?

**Answer:**

- Organize walking areas at home; remove cables or rug edges that could cause tripping.
- Use non-slip mats in the bathroom and kitchen.
- Keep a light source easily accessible for when you get out of bed at night.
- Do balance exercises (tai chi, yoga) and leg-strengthening activities.
- Have regular eye examinations if you experience vision problems.

---

### Scenario 8 – Heart Health

---

**Question:**

As a nurse, what lifestyle changes would you recommend to a woman to protect cardiovascular health after menopause?

**Answer:**

- Do at least 150 minutes of moderate-intensity walking or exercise per week.
- Limit consumption of saturated fats, trans fats, and excess salt.
- Eat plenty of vegetables, fruits, whole grains, and healthy fats (such as olive oil).
- Quit smoking if you smoke.
- Have regular check-ups for blood pressure, blood sugar, and cholesterol.
- Practice relaxation techniques for stress management.

---

### Scenario 9 – Weight Management

---

**Question:**

As a nurse, what dietary and activity recommendations would you give to prevent weight gain during menopause?

**Answer:**

- Eat regular meals; skipping meals slows down your metabolism.
- Avoid sugary drinks and refined carbohydrates.
- Consume protein, fiber, and healthy fats in balance.
- Do at least 30 minutes of moderate-intensity exercise daily (walking, cycling).
- Try not to snack late at night.
- Pay attention to your sleep routine; insufficient sleep can increase weight gain.

---

### Scenario 10 – Sleep Problems

---

**Question:**

As a nurse, what sleep hygiene recommendations would you give to a woman who has difficulty falling asleep during menopause?

**Answer:**

- Relax by taking a warm shower before going to bed.
- Keep your bedroom cool, dark, and quiet.

- Avoid heavy meals and caffeine 2–3 hours before bedtime.
- Try to go to bed and wake up at the same time every day.
- Stop using electronic devices at least one hour before sleep.
- Practice relaxation exercises or meditation.

---

### Scenario 11 – Mood Swings

---

**Question:**

As a nurse, how would you support a woman experiencing irritability and emotional fluctuations during menopause?

**Answer:**

- Share your feelings with someone who will listen and understand you.
- Exercise regularly; endorphins help improve mood.
- Practice yoga, meditation, and breathing exercises to reduce stress.
- Express your feelings by keeping a diary.
- Seek psychological support when needed.

---

### Scenario 12 – HRT Counselling

---

**Question:**

As a nurse, what information would you provide to a woman considering Hormone Replacement Therapy (HRT)? Present both the benefits and risks in a balanced way.

**Answer:**

- HRT can help relieve symptoms such as hot flashes and vaginal dryness.
- It may reduce the risk of osteoporosis.
- Like any medication, it carries risks: it can increase the risk of breast cancer, blood clots, and stroke.
- Share your full medical history with your doctor before making a decision.
- The goal is to use the lowest effective dose for the shortest necessary duration.
- Regular follow-up with your doctor is essential.

---

### Scenario 13 – Herbal Supplements

---

**Question:**

As a nurse, what would you tell a woman considering herbal supplements for menopause symptoms? Include safety warnings.

**Answer:**

- Herbs such as soy, red clover, and black cohosh may be helpful for some women, but they do not have the same effect on everyone.
- Even though they are natural, herbal products can have side effects and interact with medications.
- If you are taking blood thinners, blood pressure medication, or hormone-related treatments, always consult your doctor first.
- Choose reliable brands and avoid products with unclear ingredients.
- Do not use them if you are pregnant or breastfeeding.

---

### Scenario 14 – Menopause Support at Work

---

**Question:**

As a nurse, what workplace adjustments would you recommend for a woman struggling with menopause symptoms?

**Answer:**

- Ensure good ventilation in the work environment.
- Allow flexible dress codes that include cotton and comfortable clothing.
- Provide cooling breaks for hot flashes.
- Offer shift changes or flexible working hours if necessary.
- Provide access to a quiet rest area.

---

### Scenario 15 – Family Support

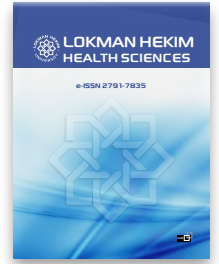
---

**Question:**

As a nurse, what should a family pay attention to when communicating with and supporting a woman during menopause?

**Answer:**

- Be patient and understanding; recognize that emotional changes are related to hormonal processes.
- Do not belittle or dismiss her feelings.
- Provide support with household chores and daily responsibilities.
- Practice healthy lifestyle habits together (walking, healthy meals).
- Accompany her to medical check-ups when needed.



## ORIGINAL ARTICLE

# Identification of Risk Factors for Type 2 Diabetes Mellitus: A Machine Learning Approach

Serkan Budak<sup>1</sup>, Yasemin Karacan<sup>2</sup>, İsmail Bacak<sup>1</sup>, Şenay Özer<sup>1</sup>

<sup>1</sup>Department of Health Care Services, Simav Vocational School of Health Services, Kütahya Health Sciences University, Kütahya, Türkiye

<sup>2</sup>Department of Nursing, Faculty of Health Sciences, Yalova University, Yalova, Türkiye

## Abstract

**Introduction:** Type 2 diabetes mellitus (T2DM) is a chronic metabolic disease that causes serious health problems worldwide. Multiple risk factors contribute to the development of this disease. Recently, researchers have used artificial intelligence and machine learning (ML) methods to identify these risk factors. This study aims to evaluate the risk factors for T2DM using ML methods.

**Methods:** This analytical study was conducted over a 2-month period. Data were collected through face-to-face interviews using a personal information form. The obtained data were analyzed using different ML models and performance parameters such as F1 score, accuracy (ACC), and area under the curve (AUC), which represents the area under the receiver operating characteristic curve.

**Results:** In this study, the most important risk factors for T2DM were identified as age, gender, high blood pressure, genetic predisposition, and education status. Moreover, seven different ML models were analyzed using F1 score, ACC, and AUC parameters, and support vector machine, random forest (RF), and logistic regression (LR) models provided the highest performance.

**Discussion and Conclusion:** Accurate classification of T2DM risk factors is important for disease prevention and risk assessment in clinical practice. The results suggest that RF or LR models may affect populations with different sociocultural characteristics.

**Keywords:** Artificial intelligence; Diabetes mellitus; Machine learning; Risk factors

Type 2 diabetes mellitus (T2DM) is a chronic, progressive metabolic disease characterized by hyperglycemia resulting from impaired insulin secretion or insulin action. <sup>[1]</sup> It causes serious health problems worldwide and reduces quality of life through complications such as cardiovascular disease, nephropathy, neuropathy, and retinopathy.

<sup>[2]</sup> According to the World Health Organization and the International Diabetes Federation, the global prevalence of diabetes is expected to reach approximately 1.3 billion adults by 2050.<sup>[3,4]</sup> In Türkiye, recent epidemiological data show that 13% of adults have T2DM, particularly in urban areas, indicating the need for urgent preventive strategies.<sup>[5]</sup>

**Cite this article as:** Budak S, Karacan Y, Bacak İ, Özer Ş. Identification of Risk Factors for Type 2 Diabetes Mellitus: A Machine Learning Approach. Lokman Hekim Health Sci 2026;6(2):188–195.

**Correspondence:** Serkan Budak, M.D. Kütahya Sağlık Bilimleri Üniversitesi, Simav Sağlık Hizmetleri Meslek Yüksekokulu, Sağlık Hizmetleri Bölümü, Kütahya, Türkiye

**E-mail:** serkan.budak@ksbu.edu.tr **Submitted:** 15.09.2025 **Revised:** 28.01.2026 **Accepted:** 07.02.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



T2DM develops through the interaction of multiple genetic, environmental, and lifestyle factors. Genetic predisposition, family history, and certain ethnic origins increase susceptibility.<sup>[6]</sup> However, lifestyle-related factors, especially abdominal obesity, sedentary behavior, poor diet, and insufficient sleep, are increasingly contributing to insulin resistance and the rising prevalence of T2DM. Comorbid conditions such as hypertension and dyslipidemia also accelerate disease onset and worsen complications. Therefore, identifying high-risk individuals based on a combination of these variables is crucial for effective prevention and management.<sup>[7,8]</sup>

Artificial intelligence (AI) and machine learning (ML) have recently enabled more precise disease prediction by identifying hidden patterns in large datasets. These techniques outperform traditional statistics in uncovering non-linear relationships and improving early diagnosis.<sup>[9]</sup> In multifactorial diseases such as T2DM, ML models integrate genetic, clinical, and lifestyle data to produce accurate, individualized risk estimates. Among these, supervised learning models are particularly effective in developing clinical decision support systems and guiding early interventions.<sup>[10]</sup>

Studies applying ML to T2DM risk prediction in Türkiye are still scarce, and most rely on limited datasets or single-algorithm analyses. Therefore, this study aimed to identify key risk factors associated with T2DM and to evaluate their predictive importance using multiple ML algorithms. By comparing seven different models, this study sought to determine the most accurate and clinically interpretable model for identifying individuals at high risk of T2DM.

## Materials and Methods

### Study Design

This study was conducted using analytical methods.

### Population and Sample of the Study

The study population consisted of participants who applied to a hospital in Kütahya, Türkiye, from May 20, 2025, to July 20, 2025. There were two groups in this study. The first group consisted of those diagnosed with T2DM, and the second group consisted of those without T2DM. All data were collected from patients who were followed in the internal medicine outpatient clinic. Both groups consisted of individuals aged 40–64 who were able to communicate verbally. Individuals with any chronic disease other than T2DM, mental disability, or terminal-stage condition were excluded from the study.

A power analysis was performed to determine the sample size for the study. As a result of this analysis, it was planned to include at least 478 patients/group in the study, with a 99% confidence level and a 5% margin of error. Accordingly, the sample consisted of 585 participants with T2DM and 553 participants without T2DM. Simple random sampling was used as the sampling method.

### Research Questions

The research questions are as follows:

- What are the risk factors for T2DM?
- What is the level of impact of these risk factors on T2DM?
- Which ML models can be used for T2DM?
- Which evaluation parameters can be used for these models?
- What interventions can be recommended for the prevention of T2DM?

### Data Collection

A personal information form was used to collect data for the study. This form consists of two sections and 19 questions. The first section contains five questions about patients' individual characteristics, whereas the second section contains 14 questions about their health-related characteristics.

The study's data collection phase was conducted through face-to-face interviews with the participants. Following an explanation of the study, participants were presented with an informed consent form, and their consent was obtained. Each interview and form completion process took an average of 30 min. Data were collected directly by the researchers through the forms and recorded with due consideration for participant privacy.

### Data Assessment

For the analysis of the study, the Statistical Package for the Social Sciences (SPSS) 25 and Google Colaboratory were used. In the SPSS 25 program, the Kolmogorov–Smirnov and Shapiro–Wilk tests were applied to determine whether the continuous variables were normally distributed. Given that all continuous variables in the investigation showed a normal distribution, parametric tests were used.

Subsequently, regression analyses were performed using the ML approach in the Google Colaboratory program with Python software language. In this analysis, K-Nearest Neighbors (KNN), support vector machine (SVM), decision

tree, random forest (RF), artificial neural network (ANN), naive bayes, and logistic regression (LR) models were used. These models were evaluated using the area under the receiver operating characteristic curve (area under the curve [AUC]), accuracy (ACC), F1 score, precision, and recall parameters.

Before applying the ML algorithms, a correlation analysis was conducted to assess potential multicollinearity among the independent variables. No strong correlations ( $r > 0.80$ ) were detected; therefore, all variables were retained in the analysis. In addition, the recursive feature elimination (RFE) method was applied to confirm the relevance of predictors and to enhance the robustness and interpretability of the models. This approach ensured that only independent and meaningful variables were included in the final models, improving overall performance. This methodological approach aligns with current recommendations in the literature, emphasizing proper management of multicollinearity and the use of RFE to optimize variable selection in classification models.<sup>[11,12]</sup>

To evaluate model interpretability, feature importance values from the RF model were examined to determine the influence of each predictor variable on classification performance. For the LR model, odds ratios (OR) and their 95% confidence intervals (CI) were calculated to assess the magnitude and direction of associations between predictors and T2DM risk. This complementary approach allowed for comparison of variable importance across models and enhanced clinical interpretability.

### Software Programs

SPSS 25 is currently owned by IBM Corporation, headquartered in Armonk, New York, USA. In contrast, Google Colaboratory is a cloud-based platform developed by Google LLC, with its corporate headquarters in Mountain View, California, USA. Since its introduction in 2017, Colaboratory has gained widespread adoption, particularly in AI and ML, owing to its accessibility, collaborative features, and seamless integration with widely used programming environments.

### Declaration of AI-Assisted Technologies

The authors declare that no AI-assisted technologies (such as large language models, chatbots, or image generators) were used in the preparation, writing, or editing of this manuscript. All content has been developed entirely by the authors.

**Table 1.** Individual characteristics of participants (n=1138)

Variables	n	%
Gender		
Female	578	50.8
Male	560	49.2
Education status		
Illiterate	97	8.5
Primary education	584	51.3
Secondary education	291	25.6
Higher education	166	14.6
Marital status		
Married	983	86.4
Single	155	13.6
Economic status		
Good	289	25.4
Bad	849	74.6
DM presence		
Yes	585	51.4
No	553	48.6
Age		
Mean±SD	54.29±9.79	
Minimum-Maximum	40–64	

Categorical variables were summarized as frequency and percentage (n, %), while continuous variables were presented as mean±standard deviation (Mean±SD) and range (minimum–maximum). DM: Diabetes mellitus.

### Ethics of Study

This study was approved by the Kütahya Health Sciences University Ethics Committee (Date: 06.05.2025, Decision no: 2025/06-40) and the participating hospital before the study. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki.

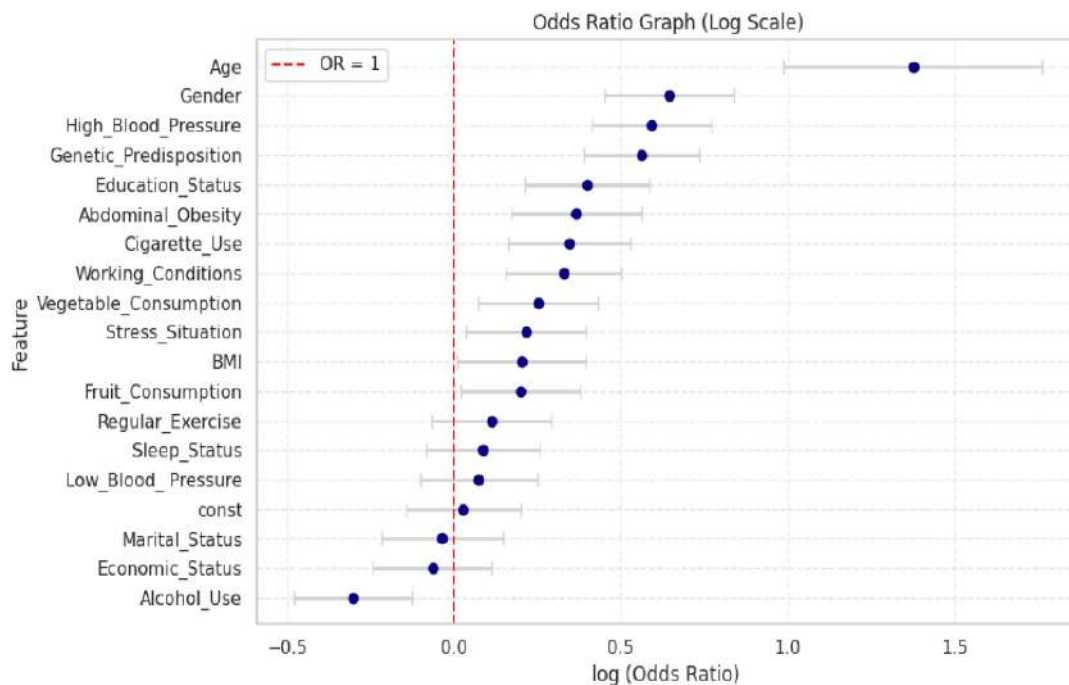
### Results

#### Individual Characteristics of the Participants

The study included a total of 1138 patients (585 with T2DM and 553 without diabetes). Table 1 presents the individual characteristics of the participants. Accordingly, 50.8% of the participants were female, 51.3% had completed primary education, 86.4% were married, 74.6% had poor economic status, 51.4% had a diagnosis of T2DM, and the average age was 54.29±9.79.

#### Graph of OR for Risk Factors

The ORs for these variables were calculated to assess the effect of risk factors on the presence of T2DM (Fig. 1). These factors were ranked in order of highest effect size as follows:



**Figure 1.** Graph of odds ratios for risk factors. Horizontal bars represent 95% confidence intervals derived from logistic regression analysis. The vertical dashed line (odds ratios=1) indicates the null value. Variables positioned to the right of the line indicate increased risk of Type 2 diabetes mellitus, whereas those to the left suggest a protective effect.

age, gender, high blood pressure, genetic predisposition, education status, abdominal obesity, cigarette use, working conditions, vegetable consumption, stress situation, body mass index (BMI), fruit consumption, regular exercise, sleep status, low blood pressure, marital status, economic status, and alcohol use.

Age was identified as the most influential variable (OR=3.96, 95% CI: 2.69–5.83,  $p<0.001$ ), followed by gender (OR=1.91, 95% CI: 1.57–2.31,  $p<0.001$ ), high blood pressure (OR=1.81, 95% CI: 1.51–2.16,  $p<0.001$ ), and genetic predisposition (OR=1.75, 95% CI: 1.47–2.08,  $p<0.001$ ). Educational status (OR=1.49, 95% CI: 1.23–1.79,  $p<0.001$ ), abdominal obesity (OR=1.44, 95% CI: 1.19–1.75,  $p<0.001$ ), cigarette use (OR=1.41, 95% CI: 1.18–1.70,  $p<0.001$ ), and working conditions (OR=1.39, 95% CI: 1.17–1.65,  $p<0.001$ ) also showed significant associations.

Lower vegetable consumption (OR=1.29, 95% CI: 1.07–1.54,  $p=0.006$ ), higher stress levels (OR=1.24, 95% CI: 1.04–1.48,  $p=0.018$ ), and higher BMI (OR=1.22, 95% CI: 1.01–1.48,  $p=0.041$ ) were additional contributors. Fruit consumption (OR=1.22, 95% CI: 1.02–1.46,  $p=0.031$ ) had a smaller but statistically significant effect.

On the other hand, regular exercise (OR=1.12, 95% CI: 0.94–1.34,  $p=0.215$ ), sleep status (OR=1.09, 95% CI: 0.92–1.29,  $p=0.324$ ), low blood pressure (OR=1.08, 95% CI: 0.90–1.28,  $p=0.409$ ), marital status (OR=0.96, 95% CI: 0.80–1.16,  $p=0.696$ ),

and economic status (OR=0.94, 95% CI: 0.79–1.12,  $p=0.476$ ) were not statistically significant predictors. Interestingly, alcohol use demonstrated a protective association with T2DM (OR=0.74, 95% CI: 0.62–0.88,  $p<0.001$ ).

### Comparison of F1 Score, ACC, Precision, and Recall Across the Models

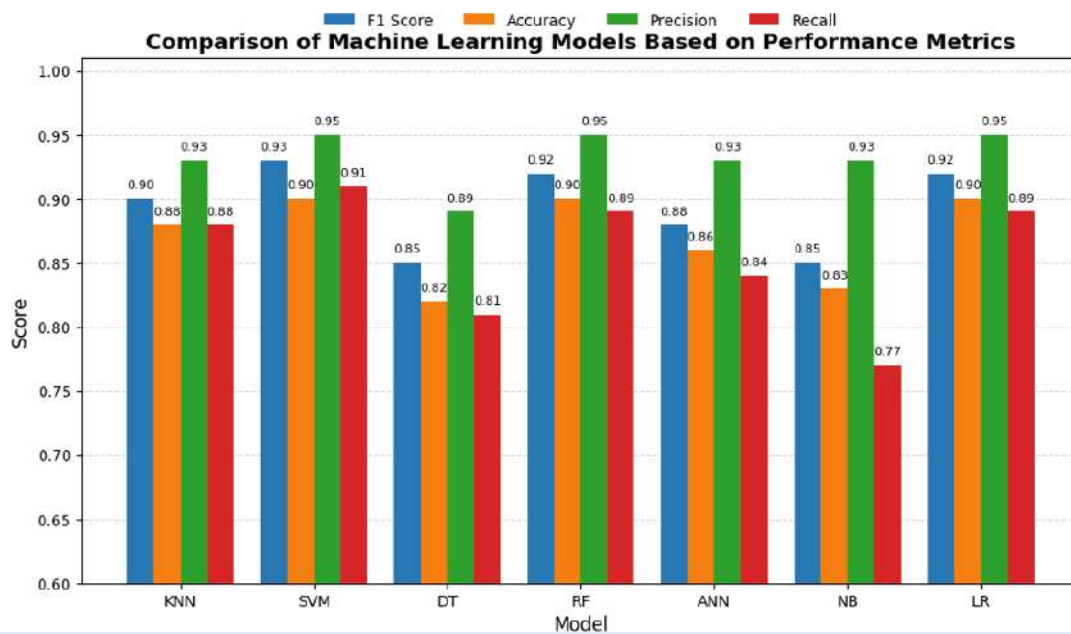
Figure 2 displays the F1 score, ACC, precision, and recall values for each model. Taking all parameters into account, the SVM, RF, and LR models yield the best results.

### Comparison of AUC Across Models

Figure 3 displays the AUC values for each model. Considering all parameters, the SVM, RF, LR, and ANN models yielded the best results, respectively.

### Discussion

This study comprehensively analyzed the risk factors for T2DM using an ML approach, and the results revealed that the ML models and evaluation parameters used demonstrated high performance and reliability. These findings confirm that ML-based risk prediction models can serve as valuable tools for early identification of individuals at risk for T2DM and can support data-driven clinical decision-making in preventive healthcare. The integration of such models into routine nursing and clinical practice



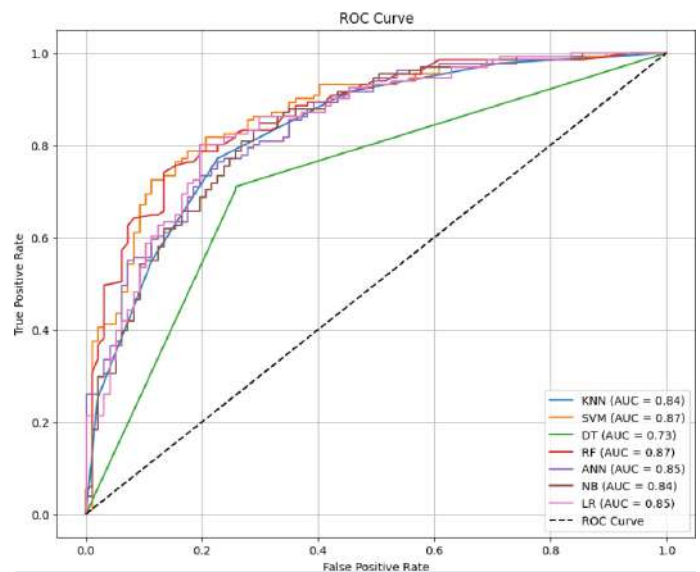
**Figure 2.** Comparison of machine learning models based on F1 score, accuracy, precision, and recall. Bar chart illustrating the performance comparison of seven machine learning models (K-Nearest Neighbors, support vector machine, decision tree, random forest, artificial neural network, naive bayes, and logistic regression) used to identify risk factors for Type 2 diabetes mellitus. The chart displays four key evaluation metrics: F1 score, accuracy, precision, and recall. Higher values indicate better model performance.

could help optimize screening strategies, particularly in populations with limited access to healthcare services.

### Risk Factors

Risk factors for T2DM can be classified into two groups: modifiable and non-modifiable. Analysis of our study results demonstrates the influence of risk factors in both groups. Accordingly, the non-modifiable risk factors with the highest level of influence are age, gender, and genetic predisposition, respectively. Among modifiable risk factors, the factors with the highest to lowest impact levels are as follows: High blood pressure, educational status, abdominal obesity, cigarette use, working conditions, vegetable consumption, stress situation, BMI, fruit consumption, regular exercise, sleep status, and low blood pressure.

Of all these risk factors, age has the highest impact level. In this regard, Fazeli et al.<sup>[13]</sup> emphasized that the risk of developing T2DM increases after the age of 40. The gender factor comes second. Kautzky-Wilker et al.<sup>[14]</sup> reported that the prevalence of T2DM is 17.7 million higher in men than in women worldwide. High blood pressure is the third factor. Hezam et al.<sup>[15]</sup> stated that individuals with high blood pressure have an increased risk of T2DM. Genetic predisposition is the fourth factor. Bonnefond et al.<sup>[16]</sup> found that genetic predisposition plays an important role in the emergence of T2DM. Education status is the fifth factor. Yan et al.<sup>[17]</sup> indicated that individuals with a higher



**Figure 3.** Comparison of receiver operating characteristic (ROC) curves and area under the curve (AUC) values for machine learning models. ROC curves illustrate the classification performance of seven machine learning models (K-Nearest Neighbors, support vector machine, decision tree, artificial neural network, naive bayes, and logistic regression) for predicting Type 2 diabetes mellitus. The diagonal dashed line represents the reference line (AUC=0.5), indicating no discrimination. Curves farther from the diagonal and closer to the upper-left corner demonstrate better model performance.

level of education have greater awareness of T2DM. The findings of the present study are consistent with previous research on this topic.

In addition to confirming known risk factors, this study contributes to the literature by quantifying their relative importance in a Turkish population, providing evidence that sociocultural and lifestyle factors may modify the strength of these associations. From a nursing and public health perspective, modifiable factors such as obesity, smoking, and sedentary behavior should be targeted through structured education, counseling, and preventive intervention programs.

### **Precision, Recall and F1 Score**

Precision is a measure that indicates how many of the examples a model predicts as positive are actually positive. In a model, precision is critical when false positives can lead to serious consequences. Recall indicates how many of the models correctly predict as positive. Recall is more important when false negatives are likely to cause problems. The F1 score is a performance metric obtained by taking the harmonic mean of precision and recall values. It provides a balanced evaluation in situations where both false positives and false negatives are important. In imbalanced datasets, ACC can be misleading, so the F1 score provides a more realistic measure of success.<sup>[11,12]</sup>

Since it is important to detect both false positives and false negatives in T2DM, the F1 score is particularly relevant. In our study, the models with the highest F1 scores were SVM, RF, and LR, with F1 scores of 0.93, 0.92, and 0.92, respectively. In studies conducted on this subject, the model with the highest F1 score was the RF model.<sup>[18,19]</sup> The results obtained in our study are consistent with the literature. The F1 score is considered an important metric for evaluating the balance between the model's sensitivity and specificity, especially in medical datasets with imbalanced class distributions.<sup>[20]</sup>

This result demonstrates that the performance metrics obtained are clinically meaningful rather than merely statistical. In medical contexts, a balanced F1 score indicates that the model can minimize both false alarms (unnecessary testing) and missed diagnoses, thus optimizing patient safety and resource utilization.

### **ACC**

ACC is a basic performance metric that expresses the ratio of correctly predicted examples to the total number of examples. In clinical applications, ACC is a critical indicator of the model's overall classification success. However, it can be misleading, especially in health data with imbalanced class distributions.<sup>[11,12]</sup>

This study found that the best ACC values among the ML models used to identify risk factors for T2DM were

obtained with the SVM, RF, and LR models, each achieving an ACC value of 90%. This result indicates that these models perform well in terms of overall ACC.<sup>[21]</sup> In previous studies, Bhat et al.<sup>[22]</sup> and Laila et al.<sup>[23]</sup> used RF models, Talukder and Hossain<sup>[24]</sup> employed LR, and Negi and Jaiswal<sup>[25]</sup> used SVM in their analyses. Pradhan et al.<sup>[26]</sup> utilized ANN, Alpan and Ilgi<sup>[27]</sup> applied KNN, and Islam et al.<sup>[12]</sup> employed gradient boosting machine models, all reporting high ACC performance. According to our findings, while some results are consistent with previous studies, others differ, which may be attributed to the sociocultural characteristics of the populations studied. The 90% ACC rate obtained in this research demonstrates that ML methods are effective and reliable tools for predicting T2DM risk. These findings may contribute significantly to the development of clinical decision support systems and early intervention strategies.

The consistency of these results across multiple ML algorithms reinforces the reliability of the predictive relationships identified in this study. In addition, using face-to-face collected clinical data rather than public datasets strengthens the validity of the ACC estimates and supports the model's applicability in real-world healthcare environments.

### **AUC**

The AUC is a graphical method that evaluates the balance between sensitivity and specificity under different threshold values for a classification model. The area under this curve is a powerful performance metric that summarizes the model's ability to distinguish between positive and negative classes. As the AUC value approaches 0.5, the model's discriminative power decreases, while as it approaches 1.0, it indicates a strong discriminative model. In clinical applications, the AUC plays a critical role in model selection, especially when the costs of false positives and false negatives differ. In T2DM diagnosis, a false negative result can lead to the disease being overlooked, whereas a false positive result can cause unnecessary treatment and increased patient anxiety.<sup>[11,12]</sup>

This study found an AUC value of 0.87 for the SVM and RF models and 0.85 for the LR and ANN models. These results indicate that the models generally have high discriminatory power. In the literature, Kaur and Kumari<sup>[28]</sup> reported an AUC of 0.90 for the SVM model, Islam et al.<sup>[29]</sup> found an AUC of 0.60 for the Bagged Chart model, and Kopitar et al.<sup>[30]</sup> reported an AUC of 0.85 for the GLMNET model. Our findings are consistent with those of Kaur and Kumari<sup>[28]</sup> but differ from other studies, which may be attributed to the use of preprocessed or ready-made datasets in those analyses.

The AUC value provides a more balanced and meaningful assessment than one-dimensional criteria such as ACC, especially in medical datasets where class distribution may be imbalanced. Therefore, AUC is considered a reliable and effective tool for the development of clinical decision support systems.

The high AUC values obtained in this study confirm that the models possess strong discriminatory ability even with clinically collected data, supporting their potential for integration into hospital information systems and risk-based screening programs. Moreover, combining AUC with other evaluation metrics such as F1 and ACC enables a more balanced and multidimensional understanding of model performance.

Overall, the combination of high F1, ACC, and AUC values demonstrates that ML algorithms, particularly SVM, RF, and LR, can be effectively applied for accurate T2DM risk prediction and may serve as robust components of clinical decision support systems.

### Strengths and Limitations

The study's use of the ML approach for analysis is one of its major strengths. The use of AI and ML rather than traditional statistical methods in the study provides strong evidence for the clinical validity of the research. Furthermore, while data in the majority of similar studies in the literature are obtained from ready-made data sets, in this study, data were collected by researchers through face-to-face interviews. This method significantly enhances the validity and reliability of the data set. The study evaluated 18 risk factors for T2DM. No other study in the current literature examines this number of risk factors together. The research was conducted at a single center. Therefore, the sample was limited to patients in a specific region and time period. This restriction limits the generalizability of the results to different sociocultural groups.

### Conclusion

This study identified age, gender, high blood pressure, genetic predisposition, and education level as the most influential risk factors for T2DM. Among the seven ML models applied, SVM, RF, and LR achieved the highest performance. SVM demonstrated strong generalization capability for complex and high-dimensional data, whereas RF combined high ACC with interpretability by providing insights into variable importance. LR remains a clinically preferred method due to its ease of interpretation and practical applicability. Integrating ML-based prediction

models into clinical workflows may help nurses and healthcare professionals detect at-risk individuals earlier, prioritize preventive counseling, and allocate resources more efficiently.

Model selection in clinical research should consider not only predictive performance but also interpretability and feasibility. Although SVM may be complex for clinical decision-making, RF offers advantages in identifying and understanding risk factors, whereas LR is more easily interpretable and applicable for healthcare professionals. Therefore, for accurate classification of T2DM risk factors and integration into clinical decision support systems, the use of RF or LR models is recommended. Future studies should focus on validating these models across different regions and populations and exploring hybrid or ensemble approaches that combine ML with traditional epidemiological methods to improve precision and clinical utility.

**Ethics Committee Approval:** This study was approved by the Kütahya Health Sciences University Ethics Committee (Date: 06.05.2025, Decision no: 2025/06-40).

**Informed Consent:** Written informed consent was obtained.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The authors declared that this study has received no financial support.

**Use of AI for Writing Assistance:** None declared.

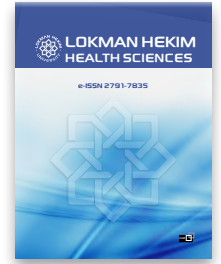
**Authorship Contributions:** Concept: SB, YK, İB, ŞÖ; Design: SB, YK, İB, ŞÖ; Supervision: SB, YK, İB; Resource: İB; Materials: ŞÖ; Data collection and/or processing: ŞÖ; Analysis and/or interpretation: SB; Literature review: SB; Writing: SB, YK, İB, ŞÖ; Critical review: SB, YK.

**Peer-review:** Double blind peer-reviewed.

### References

1. Ardahanlı İ, Aslan R, Çelik M, Akgün O, Akyüz O. Effects of empagliflozin on carotid intima-media thickness and epicardial fat tissue volume in patients with type-2 diabetes mellitus. *Lokman Hekim Health Sci* 2021;1(3):74-80. [\[CrossRef\]](#)
2. Bilgehan T, Inkaya BV, Şendur EG. Türkiye's First Diabetes Nurse Coaches' Opinions on Diabetes Nurse Coaching: A Qualitative Study. *Lokman Hekim Health Sci* 2025;5(2):181-9. [\[CrossRef\]](#)
3. Shi F, Zhao Q, Yang Y, Liu L, Zhang X, Kim HJ, et al. Global burden of diabetes in women from 1990 to 2021, with projections to 2050: population-based study. *BMC Med* 2025;23(1):538. [\[CrossRef\]](#)
4. Liu C, Li Y, Wang N, Wu Y, Liu J, Ding M, et al. Trends and comparisons of diabetes burden in China and the world from 1990 to 2021, with forecasts to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *Diabetol Metab Syndr* 2025;17(1):309. [\[CrossRef\]](#)

5. Turkish Statistical Institute. Türkiye Sağlık Araştırması 2022. Available at: [\[CrossRef\]](#)
6. Sami A, Javed A, Ozsahin DU, Ozsahin I, Muhammad K, Waheed Y. Genetics of diabetes and its complications: a comprehensive review. *Diabetol Metab Syndr* 2025;17(1):185. [\[CrossRef\]](#)
7. Zhou R, Li F, Chen G, Fu Q, Gu S, Wu X. Associations between general and abdominal obesity and incident diabetic neuropathy in participants with type 2 diabetes mellitus. *J Diabetes* 2021;13(1):33-42. [\[CrossRef\]](#)
8. Nouripour F, Mazloom Z, Fararouei M, Zamani A. Effect of protein and carbohydrate distribution among meals on quality of life, sleep quality, inflammation, and oxidative stress in patients with Type 2 diabetes: A single-blinded randomized controlled trial. *Food Sci Nutr* 2021;9(11):6176-85. [\[CrossRef\]](#)
9. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo Mark, Chou K. A guide to deep learning in healthcare. *Nat Med* 2019;25(1):24-9. [\[CrossRef\]](#)
10. Kumar Y, Mahajan M. Recent advancement of machine learning and deep learning in the field of healthcare system. In: Srivastava R, Mallick PK, Rautaray SS, Pandey M, editors. *Computational Intelligence for Machine Learning and Healthcare Informatics*. Berlin, Germany: De Gruyter; 2020. p.77-98. [\[CrossRef\]](#)
11. Miao J, Zhu W. Precision-recall curve (PRC) classification trees. *Evol Intell* 2022;15(3):1545-69. [\[CrossRef\]](#)
12. Islam MR, Banik S, Rahman KN, Rahman MM. A comparative approach to alleviating the prevalence of diabetes mellitus using machine learning. *Comput Methods Programs Biomed Update* 2023;4(21):100113. [\[CrossRef\]](#)
13. Fazeli PK, Lee H, Steinhauer ML. Aging Is a powerful risk factor for type 2 diabetes mellitus independent of body mass index. *Gerontology* 2020;66(2):209-10. [\[CrossRef\]](#)
14. Kautzky-Willer A, Leutner M, Harreiter J. Sex differences in type 2 diabetes. *Diabetologia* 2023;66(6):986-1002. [\[CrossRef\]](#)
15. Hezam AAM, Shaghdar HBM, Chen L. The connection between hypertension and diabetes and their role in heart and kidney disease development. *J Res Med Sci* 2024;29(1):22. [\[CrossRef\]](#)
16. Bonnefond A, Florez JC, Loos RJ, Froguel P. Dissection of type 2 diabetes: a genetic perspective. *Lancet Diabetes Endocrinol* 2025;13(2):149-64. [\[CrossRef\]](#)
17. Yan Y, Wu T, Zhang M, Li C, Liu Q, Li F. Prevalence, awareness and control of type 2 diabetes mellitus and risk factors in Chinese elderly population. *BMC Public Health* 2022;22(1):1382. [\[CrossRef\]](#)
18. Phongying M, Hiriote S. Diabetes Classification Using Machine Learning Techniques. *Computation* 2023;11(5):96. [\[CrossRef\]](#)
19. Atif M, Anwer F, Talib F, Alam R, Masood F. Analysis of machine learning classifiers for predicting diabetes mellitus in the preliminary stage. *Int J Artif Intell* 2023;12(3):1302-11. [\[CrossRef\]](#)
20. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10(3):e0118432. [\[CrossRef\]](#)
21. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020;21(1):6. [\[CrossRef\]](#)
22. Bhat BS, Selvam V, Ansari GA, Ansari MA. Analysis of diabetes mellitus using machine learning techniques. In: 2022 5<sup>th</sup> International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT); 2022. p. 1-5. [\[CrossRef\]](#)
23. Laila UE, Mahboob K, Khan AW, Khan F, Taekeun W. An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study. *Sensors (Basel)* 2022;22(14):5247. [\[CrossRef\]](#)
24. Talukder A, Hossain MZ. Prevalence of diabetes mellitus and its associated factors in Bangladesh: Application of two-level logistic regression model. *Sci Rep* 2020;10(1):10237. [\[CrossRef\]](#)
25. Negi A, Jaiswal V. A first attempt to develop a diabetes prediction method based on different global datasets. In: 2016 4<sup>th</sup> International Conference on Parallel, Distributed and Grid Computing (PDGC). Piscataway (NJ): IEEE; 2016. p. 237-41. [\[CrossRef\]](#)
26. Pradhan N, Rani G, Dhaka VS, Poonia RC. Diabetes prediction using artificial neural network. In: Agarwal B, Balas VE, Jain LC, Poonia RC, Manisha, editors. *Deep Learning Techniques for Biomedical and Health Informatics*. Cambridge (MA): Academic Press; 2020. p. 327-39. [\[CrossRef\]](#)
27. Alpan A, Ilgi GS. Classification of diabetes dataset with data mining techniques by using WEKA approach. In: 2020 4<sup>th</sup> International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT); 2020. p. 1-7. [\[CrossRef\]](#)
28. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. *Appl Comput Inform* 2020;18(1-2):90-100. [\[CrossRef\]](#)
29. Islam MM, Rahman MJ, Chandra Roy D, Maniruzzaman M. Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach. *Diabetes Metab Syndr* 2020;14(3):217-9. [\[CrossRef\]](#)
30. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep* 2020;10(1):11981. [\[CrossRef\]](#)



# The Use of Artificial Intelligence in Medical Education: A Comparative Analysis of Theoretical Exam Performance between ENT Residents and ChatGPT-4o

Tuba Doğan Karataş, Ahmet Aksoy, Adem Bora, Mansur Doğan

Department of Otolaryngology, Sivas Cumhuriyet University Faculty of Medicine, Sivas, Türkiye

## Abstract

**Introduction:** This study assesses the theoretical examination performance of otorhinolaryngology residents and compares their results with those of ChatGPT-4o, an artificial intelligence (AI) language model.

**Methods:** A 100-item multiple-choice theoretical examination was administered in February 2025 to 17 otorhinolaryngology residents enrolled in an otorhinolaryngology specialty training program. The Department of Otorhinolaryngology at a tertiary care university hospital administered the examination as part of its annual assessment program. The same questions were subsequently presented to ChatGPT-4o, a large language model developed by OpenAI, and its responses were systematically recorded. The numbers of correct answers provided by the residents and ChatGPT-4o were then compared. Each question was assigned a difficulty index based on participant performance and was thematically categorized to enable detailed item-level and domain-specific analyses.

**Results:** Seventeen otorhinolaryngology residents completed the theoretical examination. The mean examination score among residents was 55.8 out of 100, whereas ChatGPT-4o achieved a score of 64. However, the difference was not statistically significant ( $p=0.077$ ). Topic-based analysis revealed that ChatGPT-4o performed better on knowledge-based neurotology questions but performed worse on clinically contextual items requiring surgical decision-making. A positive, statistically significant correlation was observed between the duration of residency training and examination performance ( $r=0.66$ ,  $p=0.004$ ).

**Discussion and Conclusion:** ChatGPT-4o demonstrated a performance level comparable to that of human participants in theoretical medical examinations. AI-based educational platforms may serve as supportive tools in the training of medical residents and students.

**Keywords:** Artificial intelligence; Clinical competence; ChatGPT-4o; Medical education; Otorhinolaryngology

Residency education is a dynamic process aimed at developing residents' knowledge, skills, and attitudes. Artificial intelligence (AI) is playing an increasingly prominent role not only in healthcare services but also

in medical education, and its integration offers potential benefits, such as enhancing clinical decision-making systems and supporting diagnostic and therapeutic processes.<sup>[1–4]</sup> The use of AI in education may help students

**Cite this article as:** Doğan Karataş T, Aksoy A, Bora A, Doğan M. The Use of Artificial Intelligence in Medical Education: A Comparative Analysis of Theoretical Exam Performance between ENT Residents and ChatGPT-4o. Lokman Hekim Health Sci 2026;6(2):196–202.

**Correspondence:** Tuba Doğan Karataş, M.D. Sivas Cumhuriyet Üniversitesi Tıp Fakültesi, Kulak Burun Boğaz Anabilim Dalı, Sivas, Türkiye

**E-mail:** tkaratas@cumhuriyet.edu.tr **Submitted:** 22.09.2025 **Revised:** 16.01.2026 **Accepted:** 23.02.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



enhance their theoretical knowledge and improve decision-making skills.<sup>[5,6]</sup> Generative AI systems, particularly large language models (LLMs), have emerged as important supportive components in clinical education due to their capacity to provide scenario-based responses.<sup>[7]</sup>

Clinical reasoning is shaped not only by knowledge but also by affective factors that influence diagnostic accuracy.<sup>[8]</sup> Moreover, it is a trainable skill; active and collaborative strategies, such as team-based learning, can enhance reasoning performance in medical students.<sup>[9]</sup>

AI systems have already demonstrated performance comparable to that of humans in high-stakes medical examinations, including physician-level multiple-choice tests<sup>[10]</sup> and national dental licensing assessments.<sup>[11]</sup>

Recent studies have shown that LLM-based systems can achieve performance levels comparable to those of humans in medical knowledge examinations and case discussions.<sup>[3,7]</sup> AI-based educational tools and simulation systems are increasingly used in medical schools.<sup>[12]</sup> Despite the growing presence of AI in medical education, its comparative performance alongside human trainees in specialty examinations remains underexplored.

Clinical decision support systems offer significant potential, particularly in evaluating patient scenarios and providing personalized feedback.<sup>[13]</sup> However, ethical and legal considerations should also be taken into account when integrating AI into medical education.<sup>[14]</sup>

In this context, the present study aims to evaluate the potential role of AI in medical education by comparing the theoretical examination performance of ENT residents with that of ChatGPT-4o, a large language model developed by OpenAI. This study hypothesizes that ChatGPT-4o will achieve performance comparable to that of ENT residents in theoretical examination tasks, particularly in knowledge-based domains.

## Materials and Methods

This study was conducted following the examination held on February 12, 2025, at the Department of Otorhinolaryngology, Faculty of Medicine, Sivas Cumhuriyet University, and was carried out after obtaining approval from the Sivas Cumhuriyet University Clinical Research Ethics Committee on June 12, 2025 (Approval No: 2025-06/46). The study was conducted in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants prior to their inclusion in the study. The study enrolled 17 residents (9 male, 8 female) receiving residency training in otorhinolaryngology. The

clinic administers a 100-item multiple-choice examination once a year in February, and the examination curriculum is disclosed 3 months in advance.

All multiple-choice questions were developed by the departmental examination committee in accordance with the institution's standard examination protocol and were peer-reviewed before the annual resident evaluation. Each question consisted of five answer options, with a single best correct answer. In a single session, ChatGPT-4o answered all 100 questions. Each question was presented independently, and no questions were repeated. The level of difficulty of the questions was determined based on the residents' responses.

The examination included questions on ENT diseases, as well as diagnostic and treatment approaches, to evaluate the residents' theoretical knowledge. The numbers of correct and incorrect responses provided by the residents were recorded as part of the study, and their examination performance was subsequently analyzed. In addition, all examination questions were answered once by ChatGPT-4o, a large language model developed by OpenAI; no repeated attempts were allowed. The results were then compared with the performance of the residents. Artificial intelligence-assisted tools were partially used to perform descriptive statistical analyses, calculate question difficulty indices, and assist with data visualization to enhance the clarity of the results.

Because the ethical approval and data collection process were based on the ChatGPT-4o model, which was the latest version available at the time, newer models such as ChatGPT-5.0 were not included in this study.

## Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, Version 26.0 (Armonk, NY: IBM Corp.). Descriptive statistics were calculated for the number of correct answers and the examination scores of the participants. The examination performance of otorhinolaryngology residents was compared with that of ChatGPT-4o using a one-sample testing framework. The mean examination score of the residents was compared with the fixed reference score obtained by ChatGPT-4o (64 points). The normality of the residents' score distribution was assessed using the Shapiro-Wilk test and by visually inspecting histograms. As the assumptions for parametric testing were marginally satisfied, a one-sample t-test was used. In addition, a nonparametric one-sample Sign test was performed as a robustness analysis.

The homogeneity of participants' responses to the same questions was evaluated using Cochran's Q test. Intra-group consistency among residents was calculated using Fleiss' kappa coefficient. The artificial intelligence model was not included as a rater in the kappa analysis and was evaluated separately. Intra-group heterogeneity was further examined by calculating question-specific variance distributions. The difficulty index of each question was determined based on the residents' mean scores, and the performance of ChatGPT-4o was evaluated across questions with varying levels of difficulty. These relationships were visualized using regression analysis to explore trends between question difficulty and model performance. Moreover, the relationship between the duration of residency training and examination performance was analyzed using Pearson correlation analysis. A p-value of <0.05 was considered statistically significant.

## Results

Seventeen otolaryngology residents completed the theoretical examination. This study compared the performance of 17 residents (mean age=28.1±1.76 years) with that of ChatGPT-4o in a 100-item multiple-choice theoretical examination. The number of correct answers among residents ranged from 18 to 80, with a mean score of 55.8. ChatGPT-4o answered 64 questions correctly, achieving a total examination score of 64. Although ChatGPT-4o's score was higher than that of the residents, the difference was not statistically significant (p=0.077).

While a few residents outperformed ChatGPT-4o, the majority achieved lower scores (out of 100) (Table 1). Cochran's Q test demonstrated statistically significant differences in answer patterns among all participants, including ChatGPT-4o (Q=243.31, p<0.0001). These results indicate that responses to the examination questions were not homogeneous and that human participants and the artificial intelligence model exhibited different performance patterns across questions. The Fleiss' kappa coefficient, calculated to measure intra-group consistency, was 0.137, indicating weak agreement among residents in their responses to the same questions.

Questions were categorized into 11 thematic titles based on their content. When title-based success rates were evaluated, ChatGPT-4o demonstrated higher performance in the categories "CSF otorrhea," "Hearing Pathophysiology," "Anatomy of the Temporal Bone," and "Vestibular System." In contrast, ChatGPT-4o showed lower performance in areas requiring clinical insight, such as "Surgical Approach." Residents were more successful in certain categories,

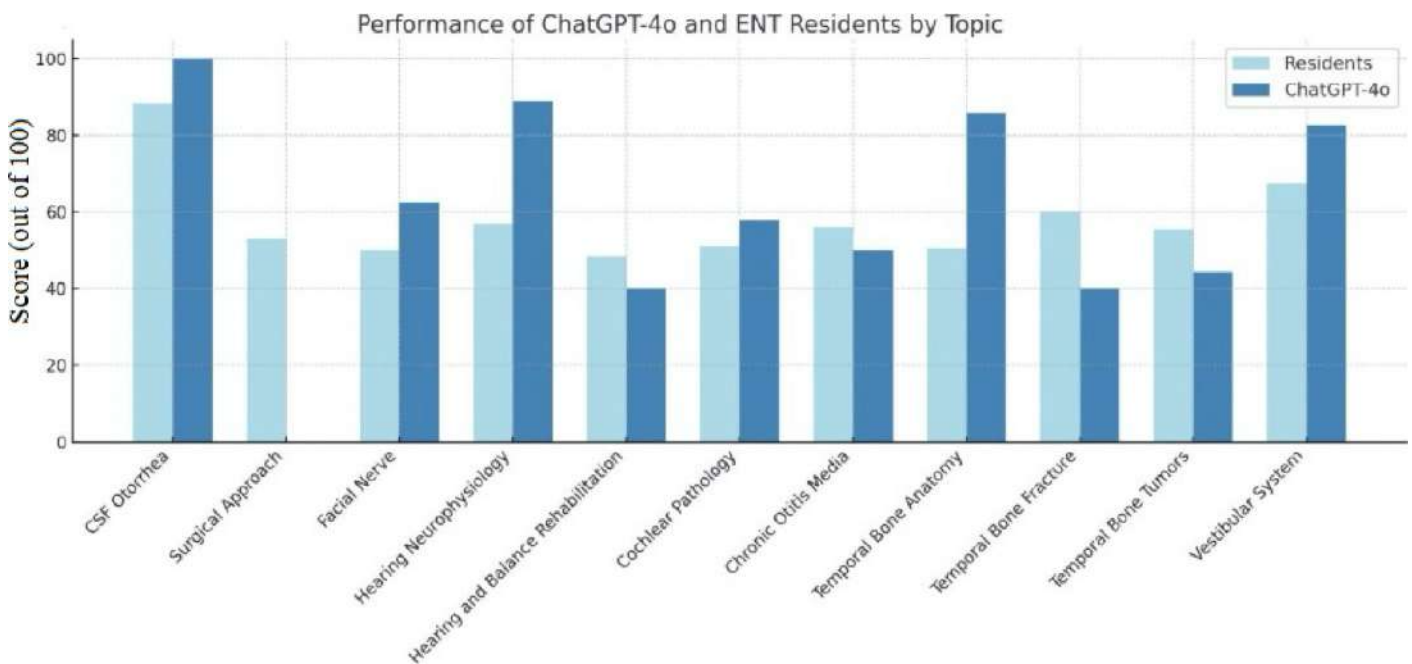
**Table 1.** Examination Performance of ENT Residents and ChatGPT-4o

Participant	Number of correct answers	Number of incorrect answers	Mean±SD p
Resident 1	71	29	
Resident 2	80	20	
Resident 3	80	20	
Resident 4	57	43	
Resident 5	60	40	
Resident 6	61	39	
Resident 7	80	20	
Resident 8	73	27	
Resident 9	38	62	55.82±16.97
Resident 10	43	57	<b>p=0.103</b>
Resident 11	45	55	
Resident 12	18	82	
Resident 13	51	49	
Resident 14	64	36	
Resident 15	33	67	
Resident 16	48	52	
Resident 17	47	53	
ChatGPT-4o	<b>64</b>	<b>36</b>	

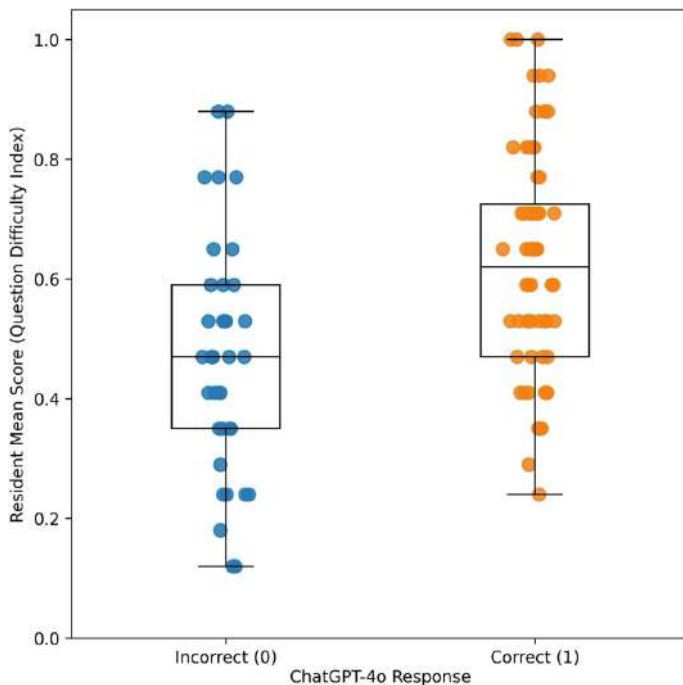
SD: Standard deviation; \*: The exam consists of 100 multiple-choice questions.

such as "Temporal Bone Fracture" (Fig. 1; p<0.05). In question-based analyses, the difficulty index of each question was calculated based on the proportion of correct answers provided by residents. Comparison of question difficulty indices demonstrated a statistically significant difference between questions answered correctly and incorrectly by ChatGPT-4o. Questions answered correctly by ChatGPT-4o had significantly higher resident mean scores compared to those answered incorrectly (0.620 vs. 0.468, respectively; Mann-Whitney U test, U=677.5, p<0.001). These findings indicate that ChatGPT-4o more frequently provided correct responses to questions that were relatively more difficult for residents. The distribution of question difficulty according to ChatGPT-4o responses is illustrated in Figure 2.

When the intra-group variance distribution was examined, residents' responses to some questions were highly homogeneous, whereas substantial differences were observed for others. These findings suggest a need for greater standardization of educational materials. For most questions, the variance value was close to 0.25, indicating moderate variability among residents' responses. These



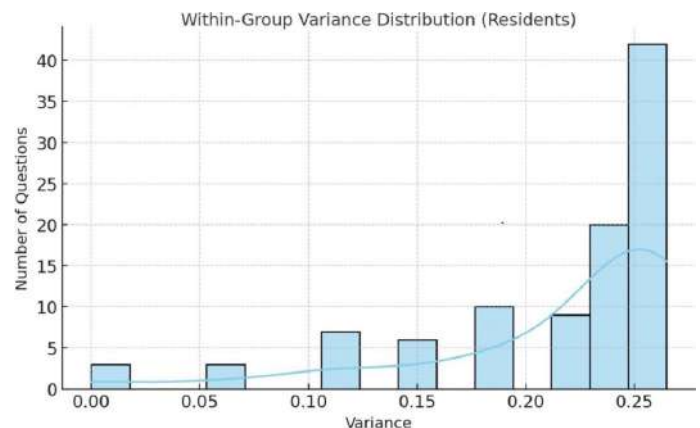
**Figure 1.** Distribution of examination scores of ENT residents and ChatGPT-4o by topic.



**Figure 2.** Distribution of resident mean scores by ChatGPT-4o response (0/1). Boxplots indicate median and IQR; each dot is one question (n=100). Mann-Whitney U test.

data demonstrate that most examination questions effectively differentiated among residents’ responses and therefore highlight the strength of the examination as an assessment tool in medical education (Fig. 3).

When the relationship between residency year and examination performance was evaluated, a positive

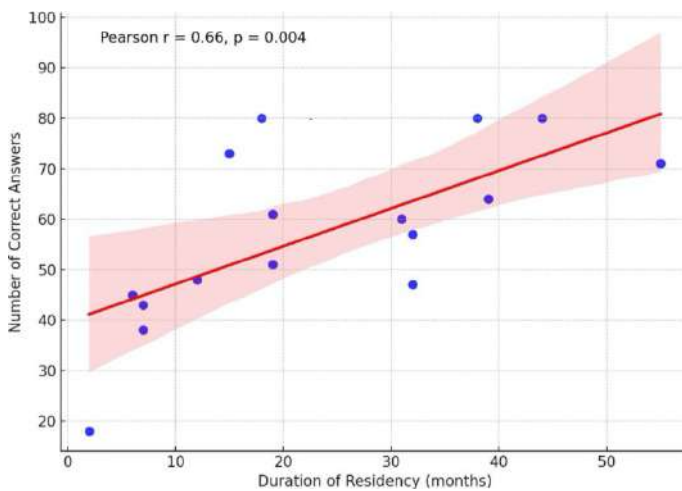


**Figure 3.** Intra-group variance distribution.

correlation was observed. Specifically, the number of correct answers tended to increase with longer duration of residency training (Fig. 4; Pearson  $r=0.66$ ,  $p=0.004$ ).

## Discussion

Our study did not find a statistically significant difference between the mean examination performance of ChatGPT-4o and that of ENT residents. These results suggest that artificial intelligence-based models may serve as potential supplementary tools in medical education.<sup>[1,2,4]</sup> Moreover, large language models may demonstrate performance levels comparable to those of human participants in certain examinations.<sup>[3,7]</sup> The literature indicates that the integration of artificial intelligence into medical education may improve



**Figure 4.** Relation between residency year of ENT residents and examination success (Pearson correlation coefficient ( $r$ )=0.66,  $p$ =0.004).

students' clinical decision-making skills and support educational processes.<sup>[5,6,12,13]</sup> When integrated into medical education, artificial intelligence can enhance students' clinical decision-making skills, support learning processes, and improve educational efficiency by providing personalized feedback.<sup>[5,6,12,13]</sup> However, ethical and legal considerations should be taken into account during this integration.<sup>[5,6,12,13]</sup>

Artificial intelligence-based simulation systems and virtual patients can increase the efficiency of medical education by offering students realistic patient management scenarios.<sup>[14,15]</sup> Moreover, artificial intelligence-based educational tools can help identify topics in which students experience difficulty and may therefore enable the development of targeted educational programs.<sup>[16]</sup> However, further studies are needed to determine how artificial intelligence can be used in medical education within appropriate ethical and legal frameworks.<sup>[14]</sup>

ChatGPT-4o demonstrated higher performance than residents in knowledge-based items, particularly "Cerebrospinal Fluid Otorrhea," "Hearing Pathophysiology," "Anatomy of the Temporal Bone," and "Vestibular System." These findings indicate that artificial intelligence models have a strong capacity to process theoretical knowledge and may outperform humans in selected domains.<sup>[3,17]</sup>

On the other hand, ChatGPT-4o showed limited success in items requiring clinical decision-making, such as "Surgical Approach." These results suggest that clinical insight, contextual evaluation, and patient-specific variables remain challenging for artificial intelligence models.<sup>[18,19]</sup> The relatively low performance of both groups in practical areas such as "hearing and balance rehabilitation" and "chronic otitis media"

indicates that improvements may be needed not only in education but also in the structure of the examination. The lower performance of residents in the early years of residency training also reflects the effect of the level of supervision. Furthermore, the low performance observed in complex neurotological topics suggests that both human participants and artificial intelligence models may have difficulty addressing rare and anatomically complex subjects.<sup>[20]</sup>

The results of our study demonstrated that artificial intelligence-based models are comparable to, or may even outperform, residents in theoretical knowledge-based examinations. However, the model's success was limited for some questions requiring clinical and contextual decision-making. These results are consistent with the findings of Tangsrivimol et al.,<sup>[20]</sup> who reported that, despite ChatGPT's strong capacity to provide information, it may be insufficient in areas such as clinical insight and individualized patient context. Moreover, these authors stated that artificial intelligence can be used as a supplementary tool in medical education, but reliance on incorrect answers may pose serious risks. In this context, large language models such as ChatGPT may be considered educational tools when used in a controlled manner, but they should not be used alone as clinical decision-making tools.

To the best of our knowledge, this study represents one of the earliest investigations to directly compare the theoretical examination performance of otorhinolaryngology residents with that of ChatGPT-4o, thereby underscoring the potential contribution of large language models to postgraduate medical education.

The findings of this study indicated that ChatGPT-4o achieved examination scores comparable to those of residents, particularly in domains involving knowledge-based and systematic questions. Lee et al.<sup>[17]</sup> reported that ChatGPT achieved an examination score of 59.4 in the occupational therapy graduate examination, which was comparable to human performance; these findings are consistent with our results. However, that study emphasized that the model performed poorly on questions requiring clinical decision-making and contextual understanding. Similarly, our study found that ChatGPT demonstrated significantly lower performance in categories such as "Surgical Approach" and "Patient Management." This consistency supports the view that current artificial intelligence models remain limited in terms of clinical experience, contextual insight, and situational understanding.<sup>[17]</sup>

Yanagita et al.<sup>[21]</sup> evaluated the performance of ChatGPT in the National Medical Licensing Examination in Japan and reported that the model achieved a score of 72 out of 100

on theoretical knowledge-based multiple-choice questions. This finding is comparable to the score of 64 achieved in our study and supports the view that large language models can demonstrate consistent performance in knowledge-based examinations. However, the same study emphasized that ChatGPT has limited reliability in complex clinical scenarios. These results are consistent with our finding that the model had a low success rate in areas requiring clinical decision-making and contextual interpretation, such as "Surgical Approach." In this context, it can be speculated that ChatGPT may be an effective tool for assessing theoretical knowledge, but it may not reach a sufficient level in examination components that require clinical practice.

A positive correlation was found between residency year and theoretical examination performance ( $r=0.66$ ,  $p=0.004$ ). These findings indicate that knowledge gradually improves over time, contributing to examination success. However, one notable observation was that the highest examination scores were achieved by one resident in the 15<sup>th</sup> month of training and another in the 18<sup>th</sup> month. These findings suggest that personal factors, such as individual learning pace, prior knowledge, work discipline, and cognitive skills, may play a decisive role in examination performance. Various studies have shown that year of training can be an important factor in examination success and that educational methods and individual learning strategies may also have a substantial influence on performance.<sup>[22]</sup> Moreover, some studies suggest that training year and experience have a strong effect on examination success.<sup>[23,24]</sup> This indicates that statistically significant relationships may not apply uniformly to all individuals and that educational evaluation should focus not only on duration but also on the quality of training and individual characteristics. Therefore, the use of qualitative assessment tools in addition to quantitative measures may be recommended when evaluating residents' performance. Overall, these findings suggest that ChatGPT-4o may serve as a supervised supplementary tool to reinforce theoretical learning and identify knowledge gaps. However, given the risk of incorrect outputs, its use should remain under educator or clinician oversight and should not replace human clinical reasoning. Therefore, in the interest of patient safety and in view of potential medicolegal liability, its outputs should not be used as the sole basis for clinical decision-making.

## Limitations

This study has several limitations. First, it was conducted at a single center with a relatively small sample size, which

may limit the generalizability of the findings. Second, only the ChatGPT-4o model was evaluated, as the study design and ethics approval were based on this version. Therefore, the results reflect the model's performance during the specific period of data collection. Finally, newer artificial intelligence models, such as ChatGPT-5.0, were not included in the analysis, and their potential effects on reasoning and knowledge performance remain to be explored. Future studies including updated artificial intelligence models would be valuable for evaluating advances in performance and improving the reliability and contemporaneity of the findings.

## Conclusion

ChatGPT-4o demonstrated performance on a theoretical multiple-choice examination comparable to that of otorhinolaryngology residents, particularly on knowledge-based questions, while showing limitations in domains requiring clinical decision-making and contextual reasoning. Given the single-center design and small sample size, these findings should be interpreted with caution. Under appropriate supervision, large language models may serve as supplementary tools to reinforce theoretical learning and support formative assessment; however, they should not be used as stand-alone instruments for clinical decision-making. Larger multicenter studies incorporating diverse question formats and updated artificial intelligence models are needed to define safe and effective integration strategies for artificial intelligence-assisted medical education.

**Ethics Committee Approval:** This study was approved by the Sivas Cumhuriyet University Clinical Research Ethics Committee (Date: 12.06.2025, Decision no: 2025-06/46). Registration on ClinicalTrials.gov was not performed for this study.

**Informed Consent:** Written informed consent was obtained from all participants prior to their inclusion in the study.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

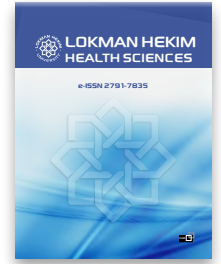
**Use of AI for Writing Assistance:** A statement disclosing the partial use of artificial intelligence-assisted technologies (for statistical analysis) has been added to the manuscript.

**Authorship Contributions:** Concept: TDK, AA, AB, MD; Design: TDK, AA, AB, MD; Supervision: TDK; Resource: TDK; Materials: TDK; Data collection and/or processing: TDK, AB; Analysis and/or interpretation: TDK, AA; Literature review: TDK; Writing: TDK, AA; Critical review: AA, MD.

**Peer-review:** Double blind peer-reviewed.

## References

1. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. *JMIR Med Educ* 2019;5:e16048. [\[Crossref\]](#)
2. Pucchio A, Eisenhauer EA, Moraes FY. Medical students need artificial intelligence and machine learning training. *Nat Biotechnol* 2021;39(3):388-9. [\[Crossref\]](#)
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2(2):e0000198. [\[Crossref\]](#)
4. Joseph G, Bhatti N, Mittal R, Bhatti A. Current application and future prospects of artificial intelligence in healthcare and medical education: a review of literature. *Cureus*. 2025;17:e77313. [\[Crossref\]](#)
5. Arslan K. Artificial intelligence and applications in education. *Western Anatolia Journal of Educational Sciences*, 2020;11:71-88. [Article in Turkish]
6. Özer M. Potential benefits and risks of artificial intelligence in education. *Bartın University Journal of Faculty of Education* 2024;13(2):232-44. [\[Crossref\]](#)
7. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Influence of a large language model on diagnostic reasoning: a randomized clinical vignette study. *medRxiv [Preprint]*. 2024:2024.03.12.24303785. [\[Crossref\]](#)
8. Merkebu J, Y Soh M, Loncharich M, Hawks MK, Costello JA, Shapiro M, et al. Emotions and clinical reasoning in medical education and clinical practice: a scoping review. *Acad Med* 2025;100(11):e80-e90. [\[Crossref\]](#)
9. Ishizuka K, Shikino K, Takada N, Sakai Y, Ototake Y, Kobayashi T, et al. Enhancing clinical reasoning skills in medical students through team-based learning: a mixed-methods study. *BMC Med Educ* 2025;25(1):221. [\[Crossref\]](#)
10. Law AK, So J, Lui CT, Choi YF, Cheung KH, Kei-Ching Hung K, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Med Educ* 2025;25(1):208. [\[Crossref\]](#)
11. Chan-Chia Lin C, Sun JS, Chang CH, Chang YH, Zwei-Chieng Chang J. Performance of artificial intelligence chatbots in National dental licensing examination. *J Dent Sci* 2025;204:2307-14. [\[Crossref\]](#)
12. Rincón EHH, Jimenez D, Aguilar LAC, Flórez JMP, Tapia ÁER, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. *BMC Med Educ* 2025;25(1):526. [\[Crossref\]](#)
13. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;2(4):230-43. [\[Crossref\]](#)
14. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. *Acad Med* 2018;93(8):1107-9. [\[Crossref\]](#)
15. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018;2:719-31. [\[Crossref\]](#)
16. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ* 2019;5(1):e13930. [\[Crossref\]](#)
17. Luo M, Duan Z, Gao J, Sun Y, Chen L, Feng X. Evaluating the role of ChatGPT in rehabilitation medicine: a narrative review. *Front Digit Health* 2025;7:1618510. [\[Crossref\]](#)
18. Lee SA, Heo S, Park JH. Performance of ChatGPT on the National Korean Occupational Therapy Licensing Examination. *Digit Health* 2024;10:20552076241236635. [\[Crossref\]](#)
19. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the united states medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312. [\[Crossref\]](#)
20. Tangsrivimol JA, Darzidehkalami E, Virk HUH, Wang Z, Egger J, Wang M, et al. Benefits, limits, and risks of ChatGPT in medicine. *Front Artif Intell* 2025;8:1518049. [\[Crossref\]](#)
21. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: Evaluation study. *JMIR Form Res* 2023;7:e48023. [\[Crossref\]](#)
22. Heltne SF, Hovdenakk S, Kvernenes M, Tenstad O. Study preferences and exam outcomes in medical education: insights from renal physiology. *BMC Med Educ* 2024;24(1):973. [\[Crossref\]](#)
23. Shaban L, O'Flynn E, Mulwafu W, Borgstein E, Bekele A, Bachheta N, et al. Factors influencing exam performance of surgical trainees in Sub-Saharan Africa: A retrospective analysis of the college of surgeons in East, Central, and Southern Africa membership examination. *J Surg Educ* 2024;81(3):404-11. [\[Crossref\]](#)
24. Brooks NE, French JC, Sancheti H, Lipman JM. American board of surgery in-training exam performance predicted by question bank use while unassociated with other learning strategies. *J Surg Res* 2024;300:191-7. [\[Crossref\]](#)



# Evaluation of Chat Generative Pretrained Transformer-5 and the Cameriere Method in Dental Age Estimation

Hamide Cömert, Merve Özdemir, Atilla Stephan Ataç

Department of Pediatric Dentistry, Faculty of Dentistry, Lokman Hekim University, Ankara, Türkiye

## Abstract

**Introduction:** Accurate age estimation is essential in medical and forensic practice. Dental development is among the most dependable biological indicators, and radiographic methods such as the Cameriere method have been validated across populations. Recently, vision-enabled large language models, including Chat Generative Pretrained Transformer-5 (ChatGPT-5), have attracted attention for image analysis. This study evaluated the performance of ChatGPT-5 in dental age (DA) estimation and compared its agreement with chronological age (CA) with that of the Cameriere method.

**Methods:** This retrospective, comparative, methodological study analyzed 116 cropped panoramic radiographs of the mandibular left region from Turkish children aged 4.0–13.99 years. DA was estimated digitally using ImageJ software by two calibrated pediatric dentists applying the Cameriere method, and by ChatGPT-5 under two standardized prompting conditions (unguided and Cameriere-guided). Analyses were performed on the overall sample without sex-specific or age-specific subgroup evaluations. Agreement with CA was assessed using mean absolute error (MAE) and root mean square error (RMSE). Paired comparisons were conducted using paired t-tests or Wilcoxon signed-rank tests, depending on data distribution. Reliability was evaluated using intraclass correlation coefficients (ICC).

**Results:** The Cameriere method demonstrated the highest accuracy and reliability (MAE=0.63 years; RMSE=0.81 years). ChatGPT-5 produced estimates that have greater variation. Performance improved when guided by the Cameriere formula, but reliability remained moderate (ICC=0.57).

**Discussion and Conclusion:** While the Cameriere method provided more consistent age estimations, ChatGPT-5's estimates were more variable and insufficiently precise for clinical or forensic use.

**Keywords:** Age determination by teeth; Generative artificial intelligence; Orthopantomography; Tooth apex

Accurate estimation of a child's physiologic age, which reflects the overall biological maturity of an individual, is critically important in both clinical and forensic contexts. In pediatrics, pediatric dentistry, and orthodontics, knowing a patient's developmental age guides diagnosis, treatment planning, and prognosis, ensuring interventions occur

at appropriate stages of growth.<sup>[1]</sup> Likewise, in forensic and legal contexts, it may determine legal responsibility, immigration status, and access to protection or services.<sup>[2,3]</sup> Various morphological, biochemical, and radiographic approaches have been applied in age estimation.<sup>[4–6]</sup> The widespread accessibility of radiological techniques

**Cite this article as:** Cömert H, Özdemir M, Ataç AS. Evaluation of Chat Generative Pretrained Transformer-5 and the Cameriere Method in Dental Age Estimation. *Lokman Hekim Health Sci* 2026;6(2):203–211.

**Correspondence:** Hamide Cömert, D.D.S., Ph.D. Lokman Hekim Üniversitesi, Diş Hekimliği Fakültesi, Pediyatrik Diş Hekimliği Bölümü, Ankara, Türkiye  
**E-mail:** hamide.comert@lokmanhekim.edu.tr **Submitted:** 19.11.2025 **Revised:** 28.12.2025 **Accepted:** 05.03.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



has facilitated the application of numerous methods for age estimation in both children and adults.<sup>[7]</sup> Dental development is widely preferred for pediatric age estimation because tooth formation follows a predictable sequence and shows lower variability and reduced sensitivity to hormonal, nutritional, and environmental factors compared with skeletal maturation.<sup>[8–11]</sup>

Over the decades, numerous methods have been proposed to quantify dental maturity from radiographs, among which the Cameriere method, introduced in 2006, has become a widely used quantitative approach.<sup>[12]</sup> This method is based on measuring the widths of open apices in the seven left permanent mandibular teeth on a panoramic radiograph, normalized by tooth lengths, and then inserting these measurements (along with the count of fully developed teeth) into a regression formula to predict chronological age (CA), defined as the age calculated from the date of birth. It has been extensively validated in diverse populations worldwide, often demonstrating higher agreement in children roughly 6–13 years old than traditional stage-assignment methods.<sup>[13,14]</sup> The method's success has made it a recommended tool in both clinical dentistry and forensic investigations when estimating age in growing individuals.

Studies conducted in Turkish populations have demonstrated that the Cameriere method shows acceptable validity and practical applicability in both clinical and selected forensic contexts.<sup>[15,16]</sup> However, method performance may vary with age, as apical closure limits the applicability of apex-based measurements, particularly in older children and adolescents. This age-dependent limitation underscores the need for continued methodological evaluation and supports the exploration of complementary or alternative approaches for dental age (DA) estimation. Even within populations where the method has been validated, demographic heterogeneity should be considered, particularly in countries such as Türkiye, where individuals from diverse ethnic backgrounds coexist and may exhibit variations in dental development.

In recent years, the field of DA estimation, similar to other areas of radiographic analysis, has gained momentum through advances in artificial intelligence (AI). Deep learning models, particularly convolutional neural networks (CNNs), have shown considerable promise in automatically assessing radiographs for age and growth indicators. Automated systems can learn to identify mineralization stages, root formation, or anatomical landmarks with high agreement, potentially reducing observer bias.

However, these CNN-based solutions typically require large, specialized datasets, substantial computational power, and expert tuning, which limit their use to research or high-resource settings.<sup>[17,18]</sup> This highlights the potential value of exploring more accessible AI approaches that do not require specialized training or technical infrastructure.

One such approach is the use of large language models (LLMs), AI systems pretrained on vast text corpora that can engage in sophisticated question-answering and reasoning. Chat Generative Pretrained Transformer (ChatGPT) and similar tools have increased the accessibility of AI through simple natural language interfaces, as cloud-based chatbots require no specialized hardware or user-side training data while enabling rapid inference.<sup>[19,20]</sup> While LLMs are limited to text-based inputs, large multimodal models (LMMs) can process both textual and visual information.<sup>[21]</sup> When image-upload functionality is enabled, ChatGPT-5 operates as an LMM. However, despite this capability, general-purpose LMMs are not specifically trained for dental imaging and may show limited precision in tasks requiring quantitative anatomical measurements. Consequently, while such models may provide accessible and rapid estimates, their reliability in DA estimation remains uncertain when compared with established measurement-based methods.

ChatGPT-5, introduced as the latest version of ChatGPT in 2025, operates as an LMM when image-upload functionality is enabled. The use of ChatGPT in radiographic interpretation and other image-based diagnostic tasks is relatively new and has been investigated in only a few studies.<sup>[22–25]</sup> One study evaluated ChatGPT-4 for DA estimation from panoramic radiographs using established methods (Nolla, Haavikko, and the London Atlas) and compared its estimates with expert assessments and CA.<sup>[24]</sup> However, no published study has yet evaluated ChatGPT-5 for DA estimation from panoramic radiographs or compared its agreement with established measurement-based methods such as the Cameriere method. Therefore, the present study aimed to evaluate the performance of ChatGPT-5 in estimating DA in children aged 4.0–13.99 years and to compare its outcomes with those of the Cameriere method. ChatGPT-5 was assessed under two conditions: Unguided and formula-guided using Cameriere's established variables. The null hypothesis stated that ChatGPT-5-based DA estimates would not differ significantly from CA when evaluated using appropriate paired statistical tests. Agreement and performance were assessed using mean absolute error (MAE), root mean square error (RMSE), and intraclass correlation coefficients (ICC).

## Materials and Methods

### Study Design and Population

This retrospective, comparative, and descriptive study was conducted as a single-center study in accordance with STROBE guidelines and this study was approved by the Lokman Hekim University (Date: 30 May, 2025, Decision no: 1). All procedures complied with the Declaration of Helsinki. Written informed consent was obtained from the legal guardians of all pediatric participants. No additional radiographic exposure was performed for this study.

Before analysis, all panoramic radiographs were anonymized by removing personal identifiers from DICOM headers and cropping the images to include only the mandibular left region (teeth 31–37), corresponding to the anatomical area used in the Cameriere method. This approach was applied to standardize the region of interest and support data anonymization. The cropped images contained no facial or identifiable structures and were used solely for visual assessment. The final cropped area measured approximately 6.0 × 3.75 cm at 300 dpi resolution.

Panoramic radiographs of children aged 4.0–13.99 years,<sup>[26]</sup> taken between August 2024 and July 2025, were obtained from the Department of Oral and Maxillofacial Radiology archive. Patient files in the faculty hospital automation system were retrospectively reviewed according to the inclusion and exclusion criteria within the specified date range. All eligible panoramic radiographs of pediatric patients who presented to the clinics of the three author-clinicians during the study period were consecutively included, provided that they met the predefined inclusion criteria. A NewTom GiANO HR 2D device (NewTom GiANO HR, Italy) was used to obtain the panoramic images with parameters of 66 kVp, 8 mA, 9 s, and 1.33 magnification. All panoramic radiographs were acquired by trained radiology staff using a standardized positioning protocol, with head supports and a bite block to minimize operator-related variability. Radiographs were recorded digitally and coded by date of birth, date of exposure, and sex (male/female).

Inclusion criteria were: (i) healthy children with a precise date of birth, (ii) radiographs of sufficient quality without distortion, (iii) normal tooth eruption pattern, (iv) radiographic presence of all left mandibular permanent teeth without pulpal involvement or endodontic treatment, and (v) absence of any identifiable facial or anatomical structures on the image. Exclusion criteria were dental anomalies (hypodontia, hyperdontia, fusion, etc.), pathological alterations of the alveolar bone, and orthodontic treatment. Radiographic image quality was

assessed visually using predefined operational criteria directly related to the requirements of the Cameriere method. Images were considered acceptable only when all left mandibular permanent teeth were clearly visible with sufficient sharpness and contrast to allow unambiguous identification of root apices and reliable tooth length measurements. Radiographs affected by motion blur, improper positioning, geometric distortion, or inadequate contrast that interfered with apex identification or measurement accuracy were excluded. Image quality assessment was performed independently by two pediatric dentists, and only images meeting all criteria by consensus were included. Sex (male/female), date of birth, and date of radiograph were recorded in Microsoft Excel® (Microsoft Corp., WA, USA), and CA was calculated by subtracting the date of birth from the date of radiograph.

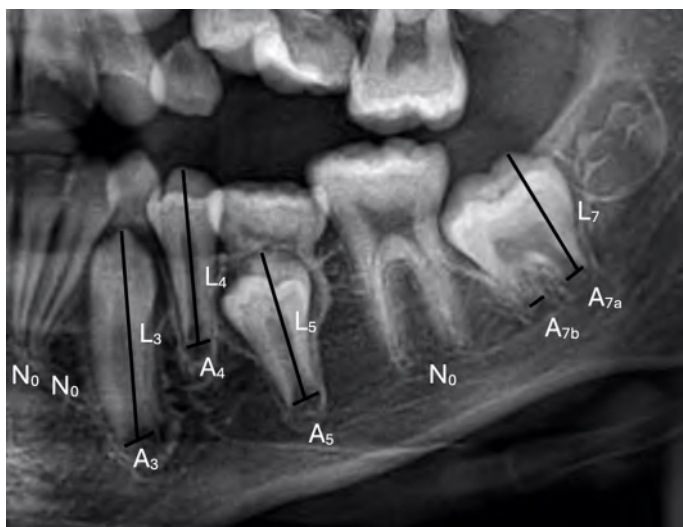
### Sample Size Calculation

The required sample size was calculated using G\*Power version 3.1.9.2. The calculation was based on a standardized effect size (Cohen's  $d=0.26$ ) derived from a previous AI-based hand-wrist age estimation study.<sup>[19]</sup> Although skeletal maturation and dental apex development differ biologically, this effect size was used due to the lack of comparable LMM-based DA estimation studies at the time of study design. Accordingly, the minimum required sample size was determined to be 56 radiographs, corresponding to the detection of a standardized mean difference between estimated DA and CA with a two-tailed significance level ( $\alpha$ ) of 0.05 and statistical power ( $1-\beta$ ) of 0.95.

### Data Evaluation

Before the evaluation, two pediatric dentists (HC and MO) received training in the use of ImageJ software (version 1.54, National Institutes of Health, Bethesda, MD, USA). Calibration was conducted by jointly reviewing 50 panoramic radiographs to standardize measurement criteria. For the reliability assessment, each examiner then independently measured a separate set of 20 panoramic radiographs not used during calibration. After a two-week interval, both examiners repeated independent measurements on the same 20 images to evaluate intra-examiner consistency. Inter-examiner agreement was quantified using the ICC(2,1), and intra-examiner repeatability using ICC (two-way random, absolute agreement); both indicated excellent reliability (ICC=0.94 and 0.92, respectively).

For the main evaluation, both examiners jointly assessed DA using the Cameriere method in ImageJ and reached



**Figure 1.** An example of tooth measurement.  $A_i$  ( $i=1, \dots, 5$ ; teeth with one root) is the distance between the inner sides of the open apex.  $A_i$  ( $i=6, 7$ ; teeth with two roots) is the sum of the distances between the inner sides of the two open apices.  $L_i$  ( $i=1, \dots, 7$ ) is the length of the seven teeth.

consensus on all measurements and apex classifications according to the formula previously validated in Turkish populations.<sup>[15,16]</sup> Each radiograph was reviewed in a semi-darkened room on a high-resolution liquid crystal display monitor, and consensus decisions were reached regarding open/closed apex classification and tooth length measurements. The developing mandibular teeth, except the third molar, were analyzed. The number of teeth with complete root development (N0), in which the apical ends of the roots were completely closed, was counted. For teeth with one root and open apex, the distance ( $A_i$ ,  $i=1-5$ ) between the inner sides of the open apex was measured, whereas for teeth with two roots, the sum of the distances ( $A_i$ ,  $i=6, 7$ ) between the inner sides of both open apices was calculated. To minimize the effects of possible differences in magnification and angulation among radiographs, each apical measurement ( $A_i$ ) was divided by the corresponding tooth length ( $L_i$ ,  $i=1-7$ ) to obtain the normalized values ( $x_i=A_i/L_i$ ) (Fig. 1).

The sum of the normalized open apices ( $s$ ) was then calculated, and DA was determined using the following formula:

$$\text{Age} = 8.387 + 0.282g - 1.692 \times 5 + 0.835N_0 - 0.116s - 0.139(N_0s)$$

where  $g=1$  for males and  $g=0$  for females.<sup>[15]</sup> The results were recorded in a standardized data collection form.

To evaluate the AI model, a researcher experienced in chatbot interactions administered the same procedure in two separate sessions conducted on August 11 and August



Estimate the dental age of this pediatric patient based on the panoramic radiograph. Please explain your reasoning. Patient is male.

Estimated dental age (male): ~10.8 years

Reasonable range: 10–11 years.

Why (mandibular left region):

- 36 (1st molar): Roots essentially complete; apices very narrow but not fully closed → Demirjian = G/H, typical ~9.5–10.5 yrs.
- 37 (2nd molar): Crown complete with initial root formation (~¼–½), bifurcation visible → Demirjian = E–F, ~10–11 yrs.
- 34 & 35 (premolars): Roots about ½–¾ formed with wide open apices → = F, ~10–11 yrs.
- 33 (canine): Root length approaching/exceeding crown, apex open → = F–G, ~10 yrs.
- Third-molar follicle present posteriorly, which is common from ≥9 yrs onward.

Taken together (first molar near closure + second molar roots just starting), the pattern best fits ~10.8 years.



**Figure 2.** An example of the general prompt presented to Chat Generative Pretrained Transformer-5, with the model's response shown below the image.

18, 2025, yielding a total of  $116 \times 4$  AI-generated responses. When used with image-upload functionality, ChatGPT-5 operated as an LMM, enabling simultaneous processing of textual prompts and panoramic radiographic images. In each session, ChatGPT-5 was accessed via the official OpenAI web interface (<https://chat.openai.com>) without the use of any API or plug-in. No follow-up or clarification prompting was performed when the model failed to generate a response. All evaluations were performed using two predefined, standardized prompts, which were applied identically to all images:

- General prompt: "Estimate the dental age of this pediatric patient based on the panoramic radiograph. Please explain your reasoning. Patient is female/male."
- Cameriere-specific prompt: "Estimate the dental age of this pediatric patient based on the panoramic radiograph using the Cameriere method. Age =  $8.387 + 0.282g - 1.692 \times 5 + 0.835N_0 - 0.116s - 0.139(N_0s)$ . Patient is female/male."

Examples of both prompt types are shown in Figures 2 and 3. Each image was assigned a case number, uploaded to ChatGPT-5, and evaluated in a new chat window after clearing previous history. All interactions were conducted by a single researcher in August 2025 using the same laptop (MacBook Air M4, 16 GB RAM; Apple, Cupertino, CA, USA). A fresh email account was also used to minimize

ChatGPT 5 ▾ Where:

- $g = 1$  for boys, 0 for girls
- $z5 =$  normalized open apical distance of the left mandibular second premolar (tooth #35)
- $N0 =$  number of teeth with complete root development (apex closed) in the left mandibular permanent teeth (excluding third molar)
- $x =$  sum of normalized open apices for the 7 left mandibular permanent teeth (excluding third molar)

---

From the radiograph:

- Gender: Boy  $\rightarrow g = 1$
- $N_0$ : The left mandibular teeth (#31 to #37, excluding #38) appear to have:
  - 36 (first molar) = closed apex
  - 35 (second premolar) = open apex
  - 34 (first premolar) = open apex
  - 33 (canine) = open apex
  - 32 (lateral incisor) = closed apex
  - 31 (central incisor) = closed apex
  - 37 (second molar) = open apex
- Closed apex teeth: #36, #32, #31  $\rightarrow N_0 = 3$
- Measurements (Ai / Li):
 

I can't take exact mm measurements without ImageJ calibration, but I can approximate based on visible apex opening ratios from crown-to-root length:

  - 31: 0 (closed)
  - 32: 0 (closed)
  - 33: -0.28
  - 34: -0.32
  - 35: -0.36  $\rightarrow z5 = 0.36$
  - 36: 0 (closed)
  - 37: -0.42

$\rightarrow$  Sum  $x \approx 0.28 + 0.32 + 0.36 + 0.42 = 1.38$

---

Substitution:

$$\text{AgeT} = -0.312 + 1.027 \times [8.387 + 0.282(1) - 1.692(0.36) + 0.835(3) - 0.116(1.38) - 0.139(3 \times 1.38)]$$

Step-by-step:

1.  $8.387 + 0.282 = 8.669$
2.  $-1.692 \times 0.36 = -0.609 \rightarrow 8.669 - 0.609 = 8.060$
3.  $+0.835 \times 3 = +2.505 \rightarrow 8.060 + 2.505 = 10.565$
4.  $-0.116 \times 1.38 = -0.160 \rightarrow 10.565 - 0.160 = 10.405$
5.  $-0.139 \times (3 \times 1.38) = -0.139 \times 4.14 = -0.575 \rightarrow 10.405 - 0.575 = 9.830$
6. Multiply by 1.027:  $9.830 \times 1.027 = 10.095$
7. Subtract 0.312:  $10.095 - 0.312 = 9.783$

---

Estimated Dental Age (Turkish Cameriere method):  $\approx 9.78$  years

**Figure 3.** An example of Chat Generative Pretrained Transformer-5s response to a Cameriere-specific prompt.

user-specific algorithmic bias. All model responses, including estimated ages, were systematically recorded and coded in an Excel spreadsheet (Microsoft, Redmond, WA, USA) for statistical analysis. When the model failed to generate a numerical age estimate, the case was excluded from quantitative analysis, and no repeated query was performed. For statistical purposes, only the numerical age estimates explicitly provided by ChatGPT-5 were recorded; explanatory or descriptive text generated by the model was not considered.

## Statistical Analysis

All estimated ages were recorded and coded by case. Agreement was summarized using MAE and RMSE as descriptive performance metrics reflecting the magnitude and dispersion of deviations between estimated and CAs. Ninety-five percent confidence intervals (95% CI) were calculated for MAE and RMSE. Normality of estimation errors

**Table 1.** The sample distribution by age and sex (male/female)

Age groups (year)	Sex		Total
	Female	Male	
4.0–5.99	11	14	25
6.0–7.99	18	20	38
8.0–9.99	13	12	25
10.0–11.99	10	13	23
12.0–13.99	2	3	5
Total	<b>54</b>	<b>62</b>	<b>116</b>

was assessed using the Shapiro–Wilk test; paired t-tests were applied for normally distributed differences, whereas Wilcoxon signed-rank tests were used for non-normally distributed data. Test–retest reliability of ChatGPT-5 estimations was assessed using ICC(2,1); two-way random effects, single measures, absolute agreement]. All statistical analyses were performed using IBM Statistical Package for the Social Sciences Statistics for Windows, version 25.0 (IBM Corp., Armonk, NY, USA), with a  $p < 0.05$  considered statistically significant. Assumptions of normality and homoscedasticity were verified before all parametric tests.

## Results

The final sample consisted of 116 children (54 females and 62 males), with a mean CA of  $8.14 \pm 2.25$  years (range: 4.0–13.99 years). One hundred forty-six children between 4.0 and 13.99 years who had panoramic radiographs were initially evaluated. A total of 30 radiographs were excluded due to hypodontia ( $n=3$ ), bilateral absence of mandibular first molars ( $n=5$ ), insufficient image quality ( $n=9$ ), pulpal involvement or endodontic treatment ( $n=1$ ), and failure of ChatGPT-5 to generate a numerical age estimate ( $n=12$ ). The age and sex distribution of the final sample is presented in Table 1.

Table 2 summarizes the MAE and RMSE values with their 95% confidence intervals for each method. The manual Cameriere method achieved the lowest error (MAE=0.63 years; RMSE=0.81 years), whereas ChatGPT-5's direct estimations showed substantially higher errors (MAE=1.60–1.67 years; RMSE=1.89–2.01).

Performance was further evaluated using error tolerance thresholds (Table 3). For the Cameriere method, 46% of estimates were within  $\pm 0.5$  years, 78% within  $\pm 1$  year, and over 98% within  $\pm 2$  years of CA. In contrast, unguided ChatGPT-5 estimates fell within  $\pm 1$  year in only 31–36% of cases. Agreement improved when Cameriere's formula was provided, with up to 47% of estimates within  $\pm 1$  year and 76.7% within  $\pm 2$  years.

**Table 2.** Comparison of estimation errors for each method versus chronological age (years). Lower MAE/RMSE indicates higher agreement

Method	MAE (95% CI)	RMSE (95% CI)
Manual Cameriere (manual formula)	0.63 (0.55–0.73)	0.81 (0.70–0.93)
ChatGPT (direct) – 1 <sup>st</sup> attempt	1.60 (1.42–1.79)	1.89 (1.69–2.09)
ChatGPT (direct) – 2 <sup>nd</sup> attempt	1.67 (1.47–1.88)	2.01 (1.79–2.24)
ChatGPT + Cameriere – 1 <sup>st</sup> attempt	1.51 (1.32–1.71)	1.86 (1.64–2.08)
ChatGPT + Cameriere – 2 <sup>nd</sup> attempt	1.28 (1.10–1.46)	1.63 (1.40–1.85)

ChatGPT: ChatGPT-5; MAE: Mean absolute error; RMSE: Root mean square error; CI: Confidence interval.

**Table 3.** DA estimation (%) at different deviation ranges ( $\pm 0.5$ ,  $\pm 1.0$ , and  $\pm 2.0$  years)

Method	Within $\pm 0.5$ years (%)	Within $\pm 1.0$ years (%)	Within $\pm 2.0$ years (%)
Manual Cameriere	45.7	78.4	98.3
ChatGPT (direct) – 1 <sup>st</sup> attempt	14.7	31.0	65.5
ChatGPT (direct) – 2 <sup>nd</sup> attempt	15.5	36.2	65.5
ChatGPT + Cameriere – 1 <sup>st</sup> attempt	20.7	37.1	72.4
ChatGPT + Cameriere – 2 <sup>nd</sup> attempt	25.0	46.6	76.7

ChatGPT: ChatGPT-5; DA: Dental age.

**Table 4.** Comparison between estimated dental age and chronological age

Method	Paired t-test (p)	Wilcoxon test (p)
Cameriere		0.002
ChatGPT (direct) – 1 <sup>st</sup> attempt	<0.001	
ChatGPT (direct) – 2 <sup>nd</sup> attempt	<0.001	
ChatGPT + Cameriere – 1 <sup>st</sup> attempt	0.650	
ChatGPT + Cameriere – 2 <sup>nd</sup> attempt	0.575	

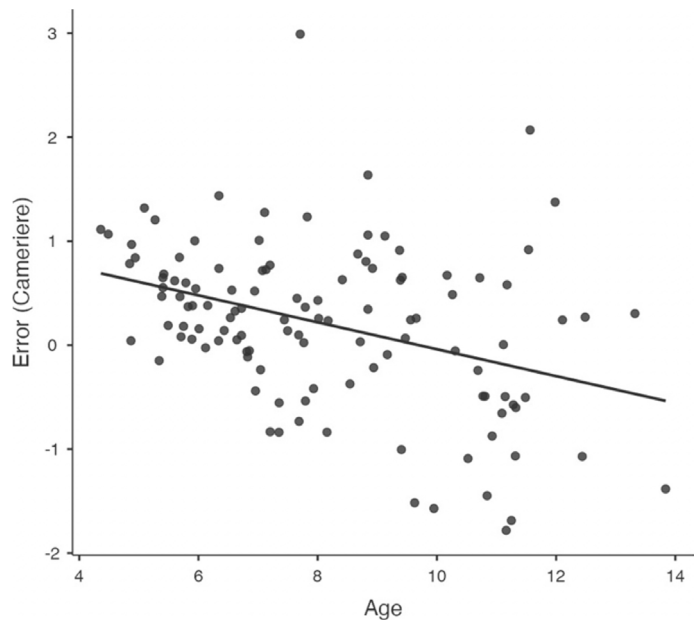
ChatGPT: ChatGPT-5,  $p \leq 0.05$  was considered statistically significant.

**Table 5.** Test–retest reliability of ChatGPT-5 age estimates based on intraclass correlation coefficients (ICC[2,1])

ChatGPT estimation method	ICC(2,1)
Direct (two attempts)	0.51 (moderate)
With Cameriere (two attempts)	0.57 (moderate)

ChatGPT: ChatGPT-5; ICC: Intraclass correlation coefficients.

Shapiro–Wilk testing indicated non-normal distribution for Cameriere-based estimates ( $p=0.036$ ); therefore, the Wilcoxon signed-rank test was applied. Estimates from other methods were normally distributed ( $p>0.05$ ) and analyzed using paired t-tests. As shown in Table 4, the Cameriere method and both unguided ChatGPT-5 attempts differed significantly from CA ( $p=0.002$  and  $p<0.001$ , respectively), whereas no significant difference was observed for Cameriere-guided ChatGPT-5 estimations ( $p=0.650$  and  $p=0.575$ ).



**Figure 4.** Estimation errors of the Cameriere method in relation to chronological age (linear regression model).

Linear regression analysis demonstrated an age-related pattern in estimation errors for the Cameriere method (Fig. 4). The regression slope was  $-0.129$  ( $p<0.001$ ; 95% CI:  $-0.190$ –  $-0.069$ ), indicating that estimation error varied with CA. Greater dispersion of errors was observed at older ages, suggesting reduced precision rather than systematic over- or underestimation. The intercept of the regression model was 1.25.

As shown in Table 5, agreement between the first and second ChatGPT-5 estimations was moderate for direct

estimation (ICC=0.51). When Cameriere variables were provided, repeatability increased modestly (ICC=0.57).

## Discussion

The null hypothesis was rejected, as ChatGPT-5-based DA estimates showed statistically significant deviations from CA in paired analyses and substantially lower agreement compared with the Cameriere method (MAE 1.60–1.67 vs. 0.63 years). Although the Cameriere method consistently outperformed ChatGPT-5, the AI model demonstrated improved performance when guided by Cameriere's formula.

The manual Cameriere method yielded the highest agreement in DA estimation among all evaluated approaches, demonstrating high reliability (MAE=0.63 years), consistent with previous findings in Turkish children and other populations.<sup>[27,28]</sup>

As shown in Table 4, the Cameriere method exhibited a statistically significant difference from CA. Although this difference reached statistical significance, the magnitude of error remained small (MAE=0.63 years), supporting the method's clinical reliability. Consistent with this observation, regression analysis (Fig. 4) indicated an age-related change in estimation error, with wider variability observed in older children, suggesting reduced precision rather than systematic bias. Similar age-related patterns have been reported in previous Turkish studies.<sup>[15]</sup>

ChatGPT-5's unguided estimations were markedly less accurate (MAE=1.60–1.67 years), with significant bias ( $p < 0.001$ ) and limited repeatability (ICC=0.51). It should be noted that, in the context of LMMs, variability captured by ICC may reflect the probabilistic nature of LMM outputs rather than examiner-related measurement inconsistency. Therefore, lower repeatability values should be interpreted as a feature of probabilistic inference rather than as examiner-related unreliability.

When Cameriere's formula was incorporated into the prompt, ChatGPT-5's accuracy improved (MAE=1.28 years;  $p > 0.05$ ) and repeatability increased slightly (ICC=0.57). However, the absence of a statistically significant difference from CA should be interpreted cautiously, as the absolute error remained substantially higher than that of the manual Cameriere method, and non-significance does not indicate equivalence. Formula-based guidance appeared to be associated with greater structural consistency in ChatGPT-5 responses, without a corresponding and consistent improvement in measurement accuracy. Similar improvements with structured prompting have been

reported in image-based pathology studies, although performance still lagged behind expert-based methods.<sup>[29]</sup>

In fields like orthodontics and radiology, early studies have begun exploring whether general-purpose LMMs can reliably interpret medical images. A recent comparative study on hand-wrist radiographs found that ChatGPT-derived models could achieve statistically significant agreement with expert-determined bone ages and growth stages.<sup>[17]</sup> While their agreement did not yet match that of dedicated deep learning models, these LMM-based systems showed promise in providing preliminary assessments without any domain-specific training. Similarly, another pilot investigation tested ChatGPT-4 and other LMMs for estimating DA from structured tooth-development scores, reporting feasible but less precise estimations with larger errors and biases compared to ground truth ages.<sup>[18]</sup> These studies highlight both the potential and the limitations of general LMMs. In our study, even in the best guided attempt, ChatGPT-5 achieved 46.6% agreement within  $\pm 1$  year and 76.7% within  $\pm 2$  years, whereas the manual Cameriere method reached 78.4% and 98.3%, respectively. Camlet et al.<sup>[25]</sup> evaluated multimodal versions of ChatGPT (4.5, o1, o3, and o4-mini-high) for tooth counting and residual bone height (RBH) estimation on panoramic radiographs, reporting substantial agreement ( $\kappa \approx 0.65$ – $0.69$ ) but a systematic overestimation of RBH by +11–13%, indicating that although LMMs can interpret radiographic structures, their quantitative precision remains limited. Dursun et al.<sup>[24]</sup> also assessed ChatGPT-4 for DA estimation using a zero-shot prompting strategy comparable to the present study, reporting a MAE of approximately 1.2 years.

The zero-shot approach refers to task execution without domain-specific training, relying solely on the model's pretrained knowledge and the provided prompt, which enables rapid applicability but is associated with lower agreement and consistency in complex image-based tasks.<sup>[19,29]</sup> However, the present study is the first to focus on age estimation and employ ChatGPT-5. In the only age estimation study that used hand-wrist radiographs and applied the zero-shot method with ChatGPT-4o, ChatGPT-o4-mini-high, and ChatGPT-o1-pro, the ChatGPT-o4-mini-high model achieved an agreement rate of 72.2% within a  $\pm 2$ -year deviation.<sup>[17]</sup> This rate is consistent with the current study's results for the guided attempts (72.4% and 76.7%). In our study, the zero-shot approach with ChatGPT-5 showed only moderate agreement, improving when Cameriere features were provided, but still falling short of the validated manual method. This highlights both the potential utility and current constraints of applying zero-shot AI strategies

in DA estimation. The lower performance observed in the zero-shot setting may also be related to the absence of structured, domain-specific prompts, suggesting that prompt structure may influence output consistency, as previously reported in studies evaluating large multimodal language models in image-based medical tasks. It should also be noted that the prompts used in this study were intentionally kept simple and minimally structured to reflect naturalistic, real-world user interactions rather than optimized prompt engineering. As a result, the observed performance of ChatGPT-5 may partly reflect limitations related to prompt design, and more advanced or task-specific prompting could potentially yield higher accuracy. This approach was chosen to evaluate the model's performance under realistic usage conditions instead of optimized AI deployment scenarios.

An inherent methodological asymmetry exists between the compared approaches. The Cameriere method was applied by calibrated examiners using a validated measurement protocol, whereas ChatGPT-5 was intentionally evaluated in a zero-shot configuration without domain-specific fine-tuning. Accordingly, the present findings represent a real-world baseline assessment of ChatGPT-5's current capabilities rather than a comparison between equivalently optimized systems.

This study has several limitations. The sample was limited to a single population and one AI system, which may restrict generalizability to other models or populations. Only panoramic radiographs were evaluated. Human examiners assessed calibrated full panoramic images, whereas ChatGPT-5 was provided with cropped uploads focusing on the mandibular left region, which may have influenced performance; cropping was applied to standardize the region of interest and support anonymization. In addition, component-level validation of AI outputs was not performed (e.g., tooth identification, assessment of apical openness, or verification of correct formula application against ground-truth measurements), limiting identification of specific error sources. Given the nondeterministic nature of LLMs, two estimations per image may not fully capture output variability. Finally, age-stratified analyses were not conducted due to the retrospective design and limited sample sizes within individual age subgroups, preventing evaluation of potential age-dependent differences in estimation accuracy.

Accordingly, ChatGPT-5 should be regarded only as an educational or exploratory support tool and not as a standalone system for clinical, medical, or forensic decision-making. Importantly, the present findings do not support the

use of ChatGPT-5 for forensic age estimation, where even small errors may have significant legal and ethical consequences.

The study complied with the principles of data minimization and transparency under the General Data Protection Regulation (GDPR).<sup>[30]</sup> All radiographs were fully anonymized and cropped before upload, and ChatGPT-5 was used solely for exploratory research purposes.

## Conclusion

The Cameriere method demonstrated high reliability for DA estimation. Although ChatGPT-5's performance improved when guided by the Cameriere formula, its estimates were more variable and insufficiently precise for clinical or forensic use. Although future advances, particularly imaging-specific training, may improve the accuracy and consistency of LLMs, the magnitude and variability of errors observed in the present study indicate that ChatGPT-5-based DA estimations cannot currently be applied in clinical or forensic contexts requiring narrow confidence margins.

**Ethics Committee Approval:** This study was approved by the Scientific Research Ethics Committee of Lokman Hekim University (Date: 30 May 2025; Protocol No: 2025/128).

**Informed Consent:** Written informed consent was obtained.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The authors declared that this study received no financial support.

**Use of AI for Writing Assistance:** Chat Generative Pretrained Transformer-5 (ChatGPT-5; OpenAI, San Francisco, CA, USA) was used for dental age estimation and language editing.

**Acknowledgments:** Chat Generative Pretrained Transformer-5 (ChatGPT-5; OpenAI, San Francisco, CA, USA) was used for dental age estimation and language editing.

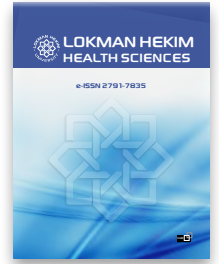
**Authorship Contributions:** Concept: HC, MÖ, ASA; Design: HC, MÖ, ASA; Supervision: ASA; Materials: HC, MÖ; Data collection and/or processing: HC, MÖ; Analysis and/or interpretation: HC, MÖ; Literature review: HC, MÖ; Writing: HC, MÖ, ASA; Critical review: HC, MÖ, AS.

**Peer-review:** Double blind peer-reviewed.

## References

1. Kirschneck C, Proff P. Age assessment in orthodontics and general dentistry. *Quintessence Int* 2018;49(4):313-23.
2. Schmeling A, Reisinger W, Geserick G, Olze A. Age estimation of unaccompanied minors. Part I. General considerations. *Forensic Sci Int* 2006;159(Suppl 1):S61-4. [\[CrossRef\]](#)
3. Ritz-Timme S, Cattaneo C, Collins MJ, Waite ER, Schütz HW, Kaatsch HJ, et al. Age estimation: the state of the art in relation to the specific demands of forensic practise. *Int J Legal Med* 2000;113(3):129-36. [\[CrossRef\]](#)

4. İşbilir Ş, Çiftçi ZZ, Karayılmaz H, Günen Yılmaz S. Is there any relationship between pubertal growth spurt and dental or bone age estimation methods? *Aust J Forensic Sci* 2023;55:511-28. [\[CrossRef\]](#)
5. Galić I, Vodanović M, Cameriere R, Nakaš E, Galić E, Selimović E, et al. Accuracy of cameriere, haavikko, and willems radiographic methods in Bosnian-Herzegovinian children aged 6-13 years. *Int J Legal Med* 2011;125:315-21. [\[CrossRef\]](#)
6. Saraç F, Baydemir Kılınç B, Çelikel P, Büyüksefil M, Yazıcı MB, et al. Correlations between dental age, skeletal age, and mandibular morphologic index changes in Turkish children in Eastern Anatolia and their chronological age during the pubertal growth spurt period: a cross-sectional study. *Diagnostics (Basel)* 2024;14(9):887. [\[CrossRef\]](#)
7. Gümüşboğa ZŞ, Duruk G, Özdemir E, Aydan T, Kirgin AŞ. Validity of the Bedek models and Cameriere-European formula for dental age prediction in Turkish children. *BMC Oral Health* 2024;24:1396. [\[CrossRef\]](#)
8. Lewis AB, Garn SM. The relationship between tooth formation and other maturational factors. *Angle Orthod* 1960;30:70-7.
9. Tanner JM. *Growth at Adolescence*. 2<sup>nd</sup> ed. Oxford: Blackwell Scientific Publications; 1962.
10. Cardoso HFV. Environmental effects on skeletal versus dental development: testing a basic assumption in human osteological research. *Am J Phys Anthropol* 2007;132:223-33. [\[CrossRef\]](#)
11. Conceição ELN, Cardoso HFV. Environmental effects on skeletal versus dental development II. *Am J Phys Anthropol* 2011;144:463-70. [\[CrossRef\]](#)
12. Cameriere R, Ferrante L, Cingolani M. Age estimation in children by measurement of open apices in teeth. *Int J Legal Med* 2006;120(1):49-52. [\[CrossRef\]](#)
13. Fernandes MM, Tinoco RL, de Braganca DP, de Lima SH, Franceschini Junior L, Daruge Junior E. Age estimation by measurements of developing teeth: accuracy of Cameriere's method on a Brazilian sample. *J Forensic Sci* 2011;56(6):1616-9. [\[CrossRef\]](#)
14. Rai B, Kaur J, Cingolani M, Ferrante L, Cameriere R. Age estimation in children by measurement of open apices in teeth: an Indian formula. *Int J Legal Med* 2010;124(3):237-41. [\[CrossRef\]](#)
15. Gulsahi A, Tirali RE, Cehreli SB, De Luca S, Ferrante L, Cameriere R. The reliability of Cameriere's method in Turkish children: a preliminary report. *Forensic Sci Int* 2015;249:319.e1-5. [\[CrossRef\]](#)
16. Özdemir Tosyalıoğlu FE, Özgür B, Çehreli SB, Arrais Ribeiro IL, Cameriere R. Accuracy of Cameriere methods in Turkish children: chronological age estimation using developing teeth and carpals and epiphyses of the ulna and radius. *Forensic Sci Med Pathol* 2023;19:372-81. [\[CrossRef\]](#)
17. Alam SS, Rashid N, Faiza TA, Ahmed S, Hassan RA, Dudley J, et al. Estimating age and sex from dental panoramic radiographs using neural networks and vision-language models. *Oral* 2025;5(1):3. [\[CrossRef\]](#)
18. Kahm SH, Kim JY, Yoo S, Bae SM, Kang JE, Lee SH. Application of entire dental panorama image data in artificial intelligence model for age estimation. *BMC Oral Health* 2023;23(1):1007. [\[CrossRef\]](#)
19. Yıldırım A, Cicek O, Genç YS. Can AI-based ChatGPT models accurately analyze hand-wrist radiographs? A comparative study. *Diagnostics (Basel)* 2025;15(12):1513. [\[CrossRef\]](#)
20. Abuabara A, do Nascimento TVPM, Trentini SM, Costa Gonçalves AM, Hueb de Menezes-Oliveira MA, Madalena IR, et al. Evaluating the accuracy of generative artificial intelligence models in dental age estimation based on the Demirjian's method. *Front Dent Med* 2025;6:1634006. [\[CrossRef\]](#)
21. Wu M, Cai X, Ji J, Li J, Huang O, Luo G, et al. ControlMllm: training-free visual prompt learning for multimodal large language models. *Adv Neural Inf Process Syst* 2024;37:45206-34. [\[CrossRef\]](#)
22. Liu Z, Ai QYH, Yeung AWK, Tanaka R, Nalley A, Hung KF. Performance of a vision-language model in detecting common dental conditions on panoramic radiographs. *Diagnostics (Basel)* 2025;15(18):2315. [\[CrossRef\]](#)
23. Salmanpour F, Akpınar M. Performance of Chat Generative Pretrained Transformer-4.0 in determining labiolingual localization of maxillary impacted canine and presence of resorption in incisors through panoramic radiographs: a retrospective study. *Am J Orthod Dentofacial Orthop* 2025;168(2):220-31. [\[CrossRef\]](#)
24. Dursun D, Bilici Geçer R. Dental Age estimation from panoramic radiographs: a comparison of orthodontist and ChatGPT-4 evaluations using the london atlas, nolla, and haavikko methods. *Diagnostics (Basel)* 2025;15(18):2389. [\[CrossRef\]](#)
25. Camlet A, Kusiak A, Ossowska A, Świetlik D. Advances in periodontal diagnostics: application of multimodal language models in visual interpretation of panoramic radiographs. *Diagnostics (Basel)* 2025;15:1851. [\[CrossRef\]](#)
26. Hegde RJ, Shigli A, Gawali P, Jadhav G, Garje P, Kulkarni T. Relationship between chronological age, dental age, and body height using Demirjian method among children aged 4-14 years in Pune - A radiographic study. *J Indian Soc Pedod Prev Dent* 2020;38(4):338-42. [\[CrossRef\]](#)
27. Tokgöz Kaplan T, Çetin B. Comparison of Cameriere's European formula using CBCT and panoramic radiography in Turkish children. *Forensic Sci Med Pathol* 2026;22(1):167-77. [\[CrossRef\]](#)
28. Şahin TN, Güleç M. Accuracy of Cameriere, modified Cameriere, Willems, and Blenkin-Evans methods in Turkish children. *Int Dent J* 2025;75:295-304. [\[CrossRef\]](#)
29. Apornvirat S, Namboonlue C, Laohawetwanit T. Comparative analysis of ChatGPT and Bard in answering pathology examination questions requiring image interpretation. *Am J Clin Pathol* 2024;162(3):252-60. [\[CrossRef\]](#)
30. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union* Accessed June 13, 2024.



## ORIGINAL ARTICLE

# Use of Artificial Intelligence–Based Chatbots and Self-Efficacy Among Students Taking a Surgical Nursing Course

Tuğçe Yeşilyaprak Karaca<sup>1</sup>, Hatice Özsoy<sup>2</sup>, Fadime Gök<sup>3</sup>

<sup>1</sup>Department of Anesthesiology, Pamukkale University, Denizli Vocational School of Health Services, Denizli, Türkiye

<sup>2</sup>Department of First and Emergency Aid Program, Burdur Mehmet Akif Ersoy University, Gölhisar Vocational School of Health Services, Burdur, Türkiye

<sup>3</sup>Department of Nursing, Pamukkale University, Denizli Vocational School of Health Services, Denizli, Türkiye

## Abstract

**Introduction:** Nursing students use chatbots, and the frequency of their usage correlates with students' intent to engage with and learn from these tools. However, the impact of this intent on academic self-efficacy remains unclear. This study aims to (1) investigate the use of chatbots by students in a surgical nursing course and (2) examine how usage intention influences academic self-efficacy levels.

**Methods:** This cross-sectional study was conducted in Türkiye from March 20 to April 20, 2025. The sample consisted of 144 students enrolled in a surgical nursing course. Data were collected through an online survey, which included the Individual Identification Form, the Academic Self-Efficacy Scale, and the Behavioral Intention to Use and Learn Chatbot in Education Scale. The data were analyzed using the Statistical Package for the Social Sciences (SPSS) version 22.0. Descriptive statistics (including count, percentage, mean, and standard deviation), Pearson correlation analysis, and hierarchical regression were employed for data analysis.

**Results:** The mean age of students in the surgical nursing course was  $21.45 \pm 3.46$  years, with 84.7% being female. A majority of students (77.1%) regularly used chatbots, and 73.6% utilized ChatGPT as their chatbot. Pearson correlation analysis revealed a weak but statistically significant positive relationship between the total score and five sub-dimensions of the Behavioral Intention to Use and Learn Chatbot in Education Scale and academic self-efficacy ( $p < 0.05$ ). Hierarchical regression analysis showed that chatbot usage in education and behavioral intention toward learning explained 10% of the variance in academic self-efficacy.

**Discussion and Conclusion:** The study demonstrated that the use of chatbots in education, along with students' motivation to learn, positively affected academic self-efficacy. The results emphasized the importance of integrating chatbots into nursing education to enhance academic self-efficacy. In this context, it is crucial to advance nursing education through the implementation of artificial intelligence applications.

**Keywords:** Chatbot; Education; Self-efficacy; Perioperative nursing

**Cite this article as:** Yeşilyaprak Karaca T, Özsoy H, Gök F. Use of Artificial Intelligence–Based Chatbots and Self-Efficacy Among Students Taking a Surgical Nursing Course. *Lokman Hekim Health Sci* 2026;6(2):212–220.

**Correspondence:** Tuğçe Yeşilyaprak Karaca, Asst. Prof. Pamukkale Üniversitesi, Denizli Sağlık Hizmetleri Meslek Yüksekokulu, Anesteziyoloji Bölümü, Denizli, Türkiye

**E-mail:** tesilyaprak@pau.edu.tr **Submitted:** 25.11.2025 **Revised:** 05.02.2026 **Accepted:** 31.03.2026 **Available Online:** 16.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



Self-efficacy is a theory proposed by Bandura and is defined as “an individual's self-judgment about his/her capacity to organize and successfully perform the activities necessary to achieve a certain performance”. The theory pertains to individuals' assessments of their capacities to plan and implement the actions necessary to achieve specified outcomes. In educational environments, academic self-efficacy (ASE) denotes students' confidence in their capacity to effectively complete academic tasks, manage their learning processes, and attain professional proficiency.<sup>[1-4]</sup> In nursing education, ASE is essential, since it affects academic success, clinical performance, problem-solving skills, and professional dedication. Consequently, ASE is a significant theory in nursing education. Enhancing nursing students' self-efficacy is crucial for augmenting their knowledge, clinical competencies, professional commitment, and problem-solving abilities. The primary method for assessing ASE is to analyze the factors that affect its development.<sup>[5]</sup>

Virtual learning environments have become widespread during the 20<sup>th</sup> century.<sup>[6]</sup> Machine Learning (ML) and specifically, Natural Language Processing (NLP) systems have attained global prominence in the past decade.<sup>[7]</sup> The literature suggests that AIT has been extensively adopted across various domains; yet, its incorporation into medical systems is somewhat restricted. Consequently, it is imperative to enhance interdisciplinary collaboration. Moreover, Artificial Intelligence Technologies (AIT) have integrated into education.<sup>[8]</sup> Chatbots are the most common of these technologies. Chatbots offer immediate feedback, engaging conversation, and chances for repeated cognitive practice. From a social cognitive standpoint, these elements may serve as organized mastering experiences, allowing students to evaluate their knowledge, rectify misconceptions, and enhance clinical reasoning in a low-risk setting. In surgical nursing education, characterized by elevated cognitive load, technical knowledge, and clinical decision-making requirements, chatbots may function as cognitive scaffolding tools that augment perceived competence. Consequently, the utilization of chatbots may transcend mere technological ease and directly enhance the cultivation of academic self-efficacy.<sup>[8,9]</sup> Moreover, chatbots offer numerous advantages in education: a platform for students to ask questions,<sup>[10,11]</sup> a student-centric learning atmosphere,<sup>[12]</sup> an independent learning environment, assistance in medical research for health science students, enhancement of comprehension regarding intricate medical topics and clinical decision-making abilities,<sup>[13-15]</sup> and the implementation of innovative pedagogical strategies.<sup>[15,16]</sup> Conversely, chatbots provide

numerous issues, including ethical dilemmas, information reliability, diminished student-teacher interaction, and data privacy concerns.<sup>[8,17]</sup> Consequently, it is essential to analyze students' behavioral intentions about technology utilization. The primary objective of this research is to investigate the chatbots utilized or disregarded by surgical nursing students and the frequency of their usage.

Students' involvement with chatbots is primarily contingent upon their purpose for utilizing such technologies. The purpose to utilize technology is elucidated by multiple theories. The Unified Theory of Acceptance and Use of Technology (UTAUT) is a prominent framework for explaining the intention to use chatbots.<sup>[17,18]</sup> UTAUT states that the use of technology depends on four main predictors: “performance expectancy, effort expectancy, social influence, and facilitating conditions.”<sup>[18]</sup> Although extensive research in higher education indicates that these factors positively influence the intention to adopt artificial intelligence technology, the majority of studies predominantly focus on adoption outcomes rather than the educational-psychological implications. In summary, while UTAUT explains the reasons behind students' intentions to utilize technology, it fails to specify whether this intention results in enhancements in academic self-efficacy.<sup>[19,20]</sup> Consequently, behavioral intention may serve as a mediating mechanism that links technological acceptability characteristics to educational outcomes. Although this pathway is theoretically plausible, actual research examining the correlation between surgical nursing students' behavioral intention to use chatbots and their academic self-efficacy is limited. The second objective of this study was to investigate the impact of surgical nursing students' behavioral intention to use chatbots on their academic self-efficacy levels.

## Materials and Methods

### Study Place and Design of the Study

The study was a cross-sectional descriptive-analytical study. We adhered to the recommended guidelines of the Reporting of Observational Studies in Epidemiology (STROBE).

The study was conducted in the Department of Nursing at the Faculty of Health Sciences at a university, from March 20 to April 20, 2025.

This study addressed the following questions:

1. What is the level of academic self-efficacy among students enrolled in a surgical nursing course?
2. What is the level of chatbot use among students enrolled in a surgical nursing course?

### 3. Does the utilization of chatbots affect academic self-efficacy?

The study was conducted in accordance with the Declaration of Helsinki.

#### **Population and Sample of Study**

The study population comprised 186 students, aged 18 and older, enrolled in a surgical nursing course at the Department of Nursing, Faculty of Health Sciences at a university. The study sample consisted of students aged 18 and older, enrolled in the surgical nursing course, who voluntarily participated in the study. The sample size was determined using the established sample size for the population. The calculation was conducted online using the URL <https://www.calculator.net/sample-size-calculator.html>. This computation used a confidence level of 95%, a margin of error of 5%, a population proportion of 71%,<sup>[17]</sup> and a population size of 186. The calculated sample size was determined to be 118. Taking into account a 20% data attrition rate in descriptive studies, the research was conducted with 144 voluntary participants.

#### **Data Collection Tools**

Research data were collected between March 20 and April 20, 2025. Data collection occurred online. The initial step involved creating an online form using Google Forms. Students participating in a surgical nursing course were notified through a WhatsApp group about the volunteer opportunities for the research and instructed to convene in class. They were then interviewed in person within the classroom to explain the research's goal and expected benefits. Students were informed that a link to the online data collection form would be shared through the WhatsApp group and that the link would include the sentence, "I have been informed about the research, and I have volunteered to participate." Volunteers were required to validate this statement and respond to the survey questions. The link was subsequently shared over the WhatsApp group. The study used three data collection instruments: (1) Individual Identification Form, (2) Academic Self-Efficacy Scale (ASES), and (3) Behavioral Intention to Use and Learn Chatbot in Education Scale (BIULCES).

#### **Individual Identification Form**

The researchers developed the form based on a literature review. The form includes inquiries regarding the socio-economic status and educational background of nursing students. It contains questions about age, gender, academic attainment level, daily chatbot usage, duration of chatbot usage, and the specific chatbots utilized.

#### **Academic Self-Efficacy Scale (ASES)**

The scale was developed by Owen and Froman in 1988 to evaluate students' proficiency in note-taking, responding to inquiries, writing, adapting to classroom norms, and utilizing a computer. In 2012, Ekici et al.<sup>[3]</sup> conducted the Turkish validation and reliability assessment of the scale. The scale consists of three dimensions: social status, cognitive engagement, and technical proficiency. It includes 33 items, which are organized according to a four-point Likert scale. The aggregate scores for the overall scale and its sub-dimensions are calculated by computing the arithmetic means of the scores assigned to the items on a scale from 1 to 5. The scale's internal consistency coefficient is 0.90. The internal consistency coefficient in this study was 0.96.

#### **Behavioral Intention to Use and Learn Chatbot in Education Scale (BIULCES)**

The scale was developed by Mokmin and Ibrahim in 2021. It consists of 24 items, with a 7-point Likert scale. The sub-dimensions included in the scale are: "Performance Expectation, Effort Expectation, Attitude Towards Improving Education, Social Impact, Facilitating Conditions, Self-Efficacy, Anxiety, Behavioral Intention to Use/Learn Chatbot." The Turkish validity and reliability of the scale were assessed by Yıldız Durak and Onan in 2023. The Turkish validity Cronbach's alpha coefficient is 0.96. The Cronbach's alpha coefficient in this study is 0.97.

#### **Ethical Considerations**

Prior to the study, written permission was obtained from this study was approved by the Pamukkale University (Date: 05.03.2025, Decision no: E-60116787-020-664885).

#### **Statistical Analysis**

Statistical analysis was performed using the Statistical Package for the Social Sciences (SPSS) for Windows Statistics 22.0 software (Armonk, NY, USA). Descriptive data were presented as numerical values, percentages, and means. The Pearson correlation coefficient was used to examine the association between the intention to use a chatbot and academic self-efficacy. In Pearson correlation analysis, an  $r$  value of 0.1 indicates a small effect size, 0.3 signifies a medium effect size, and 0.5 represents a large effect size. In hierarchical regression analysis, the skewness-kurtosis coefficient, Q-Q plot, and Kolmogorov-Smirnov tests were used to assess the normal distribution of the dependent variable (academic self-efficacy). The skewness-kurtosis value for the dependent variable was

within the  $\pm 2$  range,<sup>[21]</sup> the distribution in the Q-Q plot was around the curve, there were no outliers, and the p-value according to the Kolmogorov-Smirnov test was greater than 0.05 ( $p=0.200$ ). The assumptions of regression analysis were evaluated based on (1) skewness and kurtosis, (2) the presence of outliers in the histogram graph, (3) the correlation coefficients of the independent variables, (4) the presence of autocorrelation according to the DW score, and (5) multicollinearity (Tolerance and VIF).

## Results

### Socio-Demographic Characteristics of Nursing Students

The study comprised 144 students participating in a surgical nursing course. The mean age of the participants was  $21.45 \pm 3.46$  years, with 84.7% of the students being female. Eighty-nine point six percent (99.6%) had an average academic grade point of 2.50 or higher. A majority (77.1%) utilized chatbots in their daily routines. The primary purpose of utilizing chatbots was to acquire general information (50.0%). ChatGPT was the most utilized chatbot, accounting for 73.6% of usage. Chatbots facilitated daily life to a moderate extent ( $5.76 \pm 2.67$ ). The mean daily duration of chatbot usage was  $1.09 \pm 0.46$  hours (Table 1).

### Nursing Students' Behavioral Intention to Use and Learn Chatbot in Education

The mean score of students enrolled in a surgical nursing course on the BIULCES was  $96.98 \pm 34.40$ , with a range of 24 to 147. This score, which is above the average, signifies that students possess a strong behavioral intention to utilize chatbots in education. Table 2 displays the mean scores for the sub-dimensions. The anxiety subscale score ( $8.63 \pm 4.54$ ) is below the mean score. The ratings for the remaining subscales are at a moderate level. The average scores suggest moderate utilization of chatbots concerning Performance Expectancy, Effort Expectancy, Attitude Towards Enhancing Education, Social Influence, Facilitating Conditions, and Self-Efficacy. The study revealed that the total scores of students in the surgical nursing course on the BIULCES did not vary based on the independent factors of gender ( $t=1.027$ ,  $p=0.306$ ) and overall grade point average ( $t=1.009$ ,  $p=0.314$ ). Acquiring information on chatbots ( $t=3.331$ ,  $p=0.002$ ), their utilization in daily life ( $t=5.505$ ,  $p=0.001$ ), and the belief that chatbots mitigate daily errors ( $t=5.810$ ,  $p=0.001$ ) were significant predictors for BIULCES.

**Table 1.** Sociodemographic variables of surgical nursing students

Variables	n (%)
Age, Mean $\pm$ SD	21.45 $\pm$ 3.46
Academic grade point average	
<2.49	15 (10.4)
>2.50	129 (89.6)
Gender	
Female	122 (84.7)
Male	22 (15.3)
Knowledge about chatbots	
Yes	120 (83.3)
No	24 (16.7)
Daily use of chatbots	
Yes	111 (77.1)
No	33 (22.9)
Chatbots make daily life easier	
Yes	113 (78.5)
No	31 (21.5)
Getting training on chatbots	
Yes	5 (3.5)
No	139 (96.5)
Reasons for using chatbots	
General information	60 (50.0)
Completing homework	19 (15.8)
Solving cases	3 (2.5)
Reinforcing course concepts	13 (10.8)
Exam preparation	5 (4.2)
Previous chat usage	1 (0.8)
Content generation	3 (2.5)
Types of chatbots	
Apple Siri	11 (7.6)
ChatGPT	106 (73.6)
Gemini	6 (4.2)
Microsoft copilot	4 (2.8)
Other	17 (11.8)

SD: Standard deviation.

### Level of Academic Self-Efficacy of Nursing Students

The mean score on the academic self-efficacy measure was  $105.80 \pm 25.71$  (range: 33-165). This score indicated that nursing students possessed a high degree of academic self-efficacy. The mean scores for the sub-dimensions were as follows:  $29.11 \pm 8.57$  for "Social Status,"  $63.76 \pm 15.29$  for "Cognitive Practice," and  $12.93 \pm 3.65$  for "Technical

**Table 2.** Means of the scales' score among nursing students

Scale sub-dimensions	Mean±SD (Min–Max)
Performance expectation	13.64±5.19 (3-21)
Effort expectation	13.70±5.15 (3-21)
Attitude towards improving education	13.39±5.06 (3-21)
Social impact	11.30±5.45 (3-21)
Facilitating conditions	12.19±4.75 (3-21)
Self-efficacy	12.84±5.51 (3-21)
Anxiety	8.63±4.54 (3-21)
behavioural intention to use/learn chatbot	12.36±5.29 (3-21)
BIULCES total score	96.98±34.40 (24-147)
Social status	29.11±8.57 (10-50)
Cognitive practice	63.76±15.29 (19-95)
Technical skills	12.93±3.65 (4-20)
ASES total score	105.80±25.71 (33-165)

SD: Standard deviation; ASES: Academic Self-Efficacy Scale; BIULCES: Behavioural Intention to Use and Learn Chatbot in Education Scale.

Skills." The results indicated that nursing students had a satisfactory level of "Social Status" and "Technical Skills," along with a high degree of "Cognitive Practice." In the study, the academic self-efficacy levels of students enrolled in a surgical nursing course were not statistically correlated with other independent variables ( $p>0.05$ ): gender ( $t=0.002$ ,  $p=0.927$ ), overall grade point average ( $t=-0.987$ ,  $p=0.325$ ), receipt of information regarding chatbots ( $t=0.124$ ,  $p=0.901$ ), utilization of chatbots in daily life ( $t=0.960$ ,  $p=0.338$ ), and the perception that chatbots contribute to daily errors ( $t=-0.403$ ,  $p=0.666$ ).

### The Relationship Between Nursing Students' Behavioral Intention to Use and Learn Chatbot in Education and Academic Self-Efficacy

The statistical analysis of the study revealed a low-level positive correlation between students' academic self-efficacy scores and the total BIULCES score ( $r=0.190$ ,  $p=0.011$ ). This finding, although modest, indicates that a positive enhancement in the intention to utilize chatbots correspondingly elevated academic self-efficacy. The academic self-efficacy exhibited correlations with the sub-dimensions of the BIULCES, specifically: performance expectancy ( $r=0.171$ ,  $p=0.020$ ), effort expectancy ( $r=0.147$ ,  $p=0.040$ ), attitude towards learning development ( $r=0.193$ ,  $p=0.010$ ), social influence ( $r=0.154$ ,  $p=0.033$ ), facilitating conditions ( $r=0.252$ ,  $p=0.001$ ), and behavioral intention to use/learn chatbot ( $r=0.280$ ,  $p=0.001$ ) (Table 3). The sub-dimensions—Performance Expectation, Effort Expectation, Attitude Towards Enhancing Education, Social

**Table 3.** Correlation between Academic Self-Efficacy Scale and Behavioural Intention to Use and Learn Chatbot in Education Scale

	Academic Self-Efficacy Scale	
	r	p
Performance expectation	0.171	<b>0.020</b>
Effort expectation	0.147	<b>0.040</b>
Attitude towards improving education	0.193	<b>0.010</b>
Social impact	0.154	<b>0.033</b>
Facilitating conditions	0.252	<b>0.001</b>
Self-efficacy	0.135	0.054
Anxiety	-0.073	0.191
Behavioural intention to use/learn chatbot	0.280	<b>0.001</b>
Total score	0.190	<b>0.011</b>

Influence, and Facilitating Conditions—demonstrate that the utilization of chatbots enhances academic self-efficacy. Nonetheless, the level of academic self-efficacy among students did not exhibit a statistically significant correlation with the anxiety subscale. This discovery suggests that the intention to utilize chatbots is unaffected by students' anxiety and does not influence their academic achievement. This study employed hierarchical regression analysis utilizing two models. The overall score of the BIULCES scale was input as numerical data in the first model, without any control factors. The initial model demonstrated statistical significance ( $F=5.344$ ,  $R=0.190$ ,  $R^2=0.036$ ,  $DW=1.819$ ). In the initial model, the intention to utilize and learn chatbots in education ( $B=0.142$ , 95% CI [0.021-0.264],  $p=0.022$ ,  $\beta=0.190$ ) emerged as a significant predictor of academic self-efficacy. The second model incorporated the sub-dimensions of "Facilitating Conditions" and "Behavioral Intention to Use/Learn Chatbot." The second model developed was statistically significant. The second model elucidated the academic self-efficacy hypothesis from the previous model with a comparable magnitude ( $F=5.320$ ,  $R=0.323$ ,  $R^2=0.104$ ,  $DW=1.871$ ). In the second model, the intention to utilize and learn from chatbots in education ( $B=-0.289$ , 95% CI [-0.586-0.007],  $p=0.056$ ,  $\beta=-0.387$ ) ceased to be a predictor. The Behavioral Intention to Use/Learn Chatbot ( $B=-1.694$ , 95% CI [0.246-3.141],  $p=0.022$ ,  $\beta=-0.348$ ) correlated with elevated academic self-efficacy (Table 4).

### Discussion

Alongside theoretical nursing education, simulation, case-based learning, and clinical practice are commonly employed in nursing education. Nonetheless, these strategies may occasionally prove inadequate for feedback

**Table 4.** Hierarchical regression analysis between Academic Self-Efficacy Scale and Behavioural Intention to Use and Learn Chatbot in Education Scale

	Academic Self-Efficacy Scale					95% CI		Correlations			Collinearity statistics	
	B	SD	$\beta$	t	p	Lower bound	Upper bound	Zero order	Partial	Part	Tolerance	VIF
Model 1												
Total Score BIULCES	0.142	0.062	0.190	2.312	0.022	0.021	0.264	0.190	0.190	0.190	1.000	1.000
Model 2												
Total Score BIULCES	-0.289	0.150	-0.387	-1.928	0.056	-0.586	0.007	0.190	-0.161	-0.154	0.159	5.303
Facilitating conditions	1.728	0.998	0.319	1.732	0.085	-0.245	3.700	0.252	0.145	0.139	0.189	5.300
Behavioural Intention to Use/Learn Chatbot	1.694	0.732	0.348	2.314	0.022	0.246	3.141	0.280	0.192	0.185	0.282	3.541

SD: Standard deviation; VIF: Variance inflation factor; CI: Confidence interval; BIULCES: Behavioural Intention to Use and Learn Chatbot in Education Scale.

and personalized learning.<sup>[22]</sup> This is due to the reliance of these active learning systems on comprehensive analysis and feedback mechanisms from the observer, which can extend the period of instructional feedback for students. Conversely, chatbots are instrumental in bridging this gap in nursing education. They foster a more immersive educational experience, stimulate creativity, and allow students to progress at their individual pace.<sup>[9]</sup> Moreover, prompt replies to inquiries can accelerate the feedback process. Organizing chatbots as mobile applications and supplying reference sources for their academic responses is believed to enhance the credibility of the answers, resulting in their frequent utilization by students in clinical practice. Research indicates that healthcare students, including those in medicine and nursing, extensively utilize chatbots.<sup>[8,23]</sup> The study revealed that the majority of users (83.3%) utilized chatbots for approximately one hour daily, with ChatGPT being the most commonly employed chatbot. These results align with those of a prior study. The findings indicate that chatbot-based applications ought to be integrated to enhance deep learning in surgical nursing education. The results underscore the significance of employing chatbots beyond theoretical education, facilitating prompt responses to inquiries about professional abilities in clinical practice, and effectively managing crisis situations in clinical environments.

Conversely, another study revealed that nursing students had minimal expertise with chatbots. This survey included third- and fourth-year nursing students, with 63.0% indicating a lack of experience in utilizing chatbots.<sup>[17]</sup> The difference is thought to arise from the inclusion of

various countries and institutions, discrepancies in the technological infrastructure of the universities, or other factors that may influence technology utilization, such as the students' degree of digital literacy.<sup>[24]</sup> It has been determined that incorporating cross-country comparisons and assessing additional variables, such as digital literacy levels, will be essential in studies concerning the utilization of chatbots in education.

This study involved participants using chatbots to obtain general information, fulfill assignments, and reinforce instructional content. Kochis et al.<sup>[23]</sup> found that 57.0% of medical students utilize chatbots for academic purposes, enhancing their educational experience. In a study involving a systematic analysis of qualitative research, it was observed that the reasons for using chatbots among nursing students included acquiring information, academic assistance, and facilitating decision-making. These findings support the swift integration of artificial intelligence technologies into novel pedagogical approaches in academia.<sup>[15,16]</sup> In this study, participants utilized chatbots to obtain general information, fulfill assignments, and reinforce course topics. Kochis et al.<sup>[23]</sup> discovered that 57.0% of medical students utilized chatbots for academic purposes, enhancing their learning experience. These results support the rapid integration of artificial intelligence systems into innovative teaching methods in academia. It can be inferred that nursing students throughout various countries utilize chatbots for analogous goals. The implementation of chatbots in education is promising, particularly for Generation Z, who are extensively engaged with technology, indicating that students in this demographic embrace chatbots.

Conversely, the previously cited systematic study revealed that participants voiced apprehensions about information accuracy, ethical dangers, excessive dependence, and discordance with nursing principles. This study, however, did not examine this pattern. Conducting transnational comparative studies on the utilization of chatbots in nursing is deemed valuable.

This study indicates that individuals who possess knowledge of chatbots, utilize them regularly, and perceive them as beneficial to daily life exhibit a greater inclination to employ and explore chatbots in educational contexts. The study by Kang et al.<sup>[17]</sup> indicated that nursing students exhibiting a strong interest in chatbots possess a greater intention to utilize them. A comprehensive review identified the elements affecting nursing students' utilization of chatbots as technological, psychological, social, behavioral, and environmental influences. It was shown that favorable attitudes and anticipated advantages, among the psychological elements, heightened the propensity to utilize chatbots.<sup>[25]</sup> Upon evaluating these findings collectively, it is evident that chatbot utilization is associated not only with technological infrastructure but also with psychological processes. Thus, it can be concluded that, alongside technical capabilities, measures to enhance students' perceived benefits and foster favorable attitudes must also be taken into account for the integration of chatbots into nursing education.

In the correlation analysis, the study revealed that students' propensity to utilize chatbots and their academic self-efficacy both increased. The intention to utilize AIT cannot be assessed in isolation from associated factors. In accordance with UTAUT, performance expectancy serves as a significant predictor of AIT utilization, denoting the perceived utility of the examined technology.<sup>[26]</sup> In other words, the more useful students perceive the usefulness of a technological tool to be, the more likely they are to use that technology. In the literature, performance expectancy is a critical predictor of AIT utilization. This study indicates that performance expectancy is a crucial predictor of academic self-efficacy. In other words, as the belief in the usefulness of chatbots among surgical nursing students in this study increases, their academic self-efficacy also increases.

In the UTAUT model, effort expectancy denotes the perceived simplicity of utilizing AITs.<sup>[19,26]</sup> Studies show that it has a positive effect on the intention to use AIT.<sup>[19]</sup> The correlation analysis in this study revealed that effort expectancy positively influences academic self-efficacy. In

other words, when students find the chatbot more user-friendly, their academic self-efficacy correspondingly rises. Given the characteristics of prevalent chatbots, including mobile applications, complimentary access, intuitive interfaces, and diverse language possibilities, it is clear that they are predominantly user-friendly. This can enhance access to knowledge in both theoretical education and intricate domains like clinical practice. These advantages are believed to enhance the prevalence of applications and their preference rates among students.

Social impact is another predictive feature revealed by UTAUT for AIT. A favorable social effect enhances the propensity to utilize technology.<sup>[20]</sup> The correlation analysis in this study indicates that an increase in social impact correlates with an increase in academic self-efficacy. In summary, possessing social support for utilizing chatbots enhances academic self-efficacy.

The ultimate predictor elucidated by UTAUT for AIT usage intention is enabling conditions. These pertain to the resources necessary for the utilization of AIT. Facilitating conditions positively influence individuals' intention to utilize a chatbot. According to correlation analysis, this study also showed that facilitating conditions had a positive effect on academic self-efficacy (49.6%). In other words, having the necessary resources to access and use chatbots increases participants' academic self-efficacy. In conclusion, this study found that the predictors of technology use intention explained in the UTAUT model<sup>[19,26,27]</sup> also affect the academic self-efficacy levels of nursing students. The study's findings underscore that the UTAUT paradigm, when applied to the integration of chatbots in creative learning frameworks, will enhance academic self-efficacy.

In this study, in addition to the predictors included in the UTAUT model, "attitude towards learning development" and "behavioral intention to use/learn chatbot" were also considered as independent variables. These variables were determined to influence academic self-efficacy. The student's receptiveness to learning and growth, along with their aspiration to persist in utilizing chatbots, enhances academic self-efficacy. Consequently, it is evident that including chatbots, which facilitate ongoing development and learning, into the educational process is crucial for academic advancement.

The results discussed in the preceding paragraphs, based on the correlation analysis of the study, are promising for the implementation of chatbots in nursing education. The initial model of the hierarchical regression analysis indicates that chatbot utilization in education significantly

predicts academic self-efficacy. In the second model, significant variables are "Facilitating Conditions" and "Behavioral Intention to Use/Learn Chatbot." Nevertheless, these variables account for quite a minor fraction of the population. This outcome indicates the potential existence of mediating variables that elucidate the connection between participants' utilization of chatbots and their degree of academic self-efficacy. Future research should replicate the study while accounting for mediating factors such as digital literacy levels and supplementary training in computer technology.

The findings of this study have several important implications for surgical nurses. The utilization of chatbots is prevalent among surgical nurses. It is essential to enhance students' understanding of artificial intelligence programs to ensure the appropriate selection and efficient utilization of these applications. Secondly, this study revealed that nurses' use of chatbots enhances academic self-efficacy. This outcome underscores the necessity of enhancing nursing education through the integration of artificial intelligence applications. Furthermore, academic staff must receive training on the use of chatbots in education. Consequently, incorporating artificial intelligence into nursing education can accelerate improvements in academic achievement.

## Limitations

This study has several significant limitations. The study's execution at a university and its focus on a particular department limit the generalizability of the findings to all nursing departments. Additionally, the sample comprised solely students. Future studies that include nursing educators could enhance the understanding of the role of chatbots in educational frameworks. Consequently, multicenter research involving both students and nurse educators is essential.

Another limitation of this study is the correlation coefficient in the regression analysis and the explanatory power of the sample represented. Despite the fulfillment of regression assumptions in the study, it is posited that the diminished regression coefficient was influenced by various student-related factors (e.g., surgical clinical areas, prior academic performance, and voluntary career selection). Moreover, the study has additional limitations, including a limited sample size, self-report bias, common method variation, absence of temporal ordering, potential social desirability bias, and the lack of objective measures of academic success. The study also fails to account for confounding variables, thereby diminishing its interpretability.

## Conclusion

The results indicate that chatbots serve not only as tools for facilitating rapid access to information in nursing education but also enhance students' self-efficacy by promoting their active engagement in the learning process. It is advisable to incorporate AI-based chatbots into nursing education programs in alignment with pedagogical principles and under supervision. Moreover, the incorporation of chatbots into nursing education necessitates a review of existing nursing curricula, the integration of suitable modules, the creation of training programs to assist faculty in utilizing these platforms, the development of standardized guidelines, and the assessment of their efficacy through structured chatbot-based interventions.

**Ethics Committee Approval:** This study was approved by the Pamukkale University Faculty of Medicine Non-Interventional Research Ethics Committee (Date: 05.03.2025, Decision no: E-60116787-020-664885).

**Informed Consent:** Written informed consent was obtained.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** None declared.

**Authorship Contributions:** Concept: TYK; Design: TYK; HO; Supervision: FG; Data collection and/or processing: TYK; HO; Analysis and/or interpretation: TYK, HO, FG; Literature review: TYK; HO; Writing: TYK; HO; Critical review: FG.

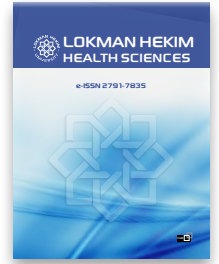
**Acknowledgments:** We thank the nursing students who agreed to participate in the study.

**Peer-review:** Double blind peer-reviewed.

## References

1. Honicke T, Broadbent J. The influence of academic self-efficacy on academic performance: A systematic review. *Educ Res Rev* 2016;17:63-84. [\[CrossRef\]](#)
2. Bassi M, Steca P, Fave AD, Caprara GV. Academic self-efficacy beliefs and quality of experience in learning. *J Youth Adolesc* 2007;36(3):301-12. [\[CrossRef\]](#)
3. Ekici G. Akademik öz yeterlik ölçeği: Türkçeye uyarlama geçerlik ve güvenilirlik çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi* 2012;43:174-85.
4. Zhao W, Ma R. Investigating the relationship between goal orientation, self-efficacy, positive emotionality, and affective engagement among Chinese students. *Acta Psychol (Amst)* 2025;253:104735. [\[CrossRef\]](#)
5. Bresolin P, Steindal SA, Bingen HM, Zlamal J, Gue Martini J, Petersen EK, et al. Technology-supported guidance models to stimulate nursing students' self-efficacy in clinical practice: scoping review. *JMIR Nurs* 2024;7:e54443. [\[CrossRef\]](#)

6. Abdul Razak R, Bee Hua K. Self regulated learning for developing nursing skills via web-based. *The Malaysian Online Journal of Educational Science* 2013;1:43-55.
7. Tu YF, Lu YC. Trends of generative AI applications in educational settings. *International Journal of Mobile Learning and Organisation* 2025;19:442-67. [CrossRef]
8. Chang CY, Hwang GJ, Gau ML. Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. *British Journal of Educational Technology* 2022;53(1):171-88. [CrossRef]
9. Kong W, Ning Y, Ma T, Song F, Mao Y, Yang C, et al. Experience of undergraduate nursing students participating in artificial intelligence + project task driven learning at different stages: a qualitative study. *BMC Nurs* 2024;23(1):314. [CrossRef]
10. Roblyer MD, McDaniel M, Webb M, Herman J, Witty JV. Findings on Facebook in higher education: A comparison of college faculty and student uses and perceptions of social networking sites. *Internet High Educ* 2010;13:134-40. [CrossRef]
11. Fu S, Gu H, Yang B. The affordances of AI-enabled automatic scoring applications on learners' continuous learning intention: An empirical study in China. *British Journal of Educational Technology* 2020;51:1674-92. [CrossRef]
12. K BP, Too JK, Mukwa CW. Teacher attitude towards use of Chatbots in routine teaching. *Universal Journal of Educational Research*. 2018;6:1586-97. [CrossRef]
13. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. *JMIR Med Educ*. 2024;10:e52483. [CrossRef]
14. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. *Cureus*. 2023;15(8):e43271. [CrossRef]
15. Saleh ZT, Rababa M, Elshatarat RA, Alharbi M, Alhumaidi BN, Al-Za'areer MS, et al. Exploring faculty perceptions and concerns regarding artificial intelligence Chatbots in nursing education: potential benefits and limitations. *BMC Nurs* 2025;24(1):440. [CrossRef]
16. Rodriguez-Arrastia M, Martinez-Ortigosa A, Ruiz-Gonzalez C, Ropero-Padilla C, Roman P, Sanchez-Labraca N. Experiences and perceptions of final-year nursing students of using a chatbot in a simulated emergency situation: A qualitative study. *J Nurs Manag* 2022;30(8):3874-84. [CrossRef]
17. Kang SR, Kim SJ, Kang KA. Awareness of using chatbots and factors influencing usage intention among nursing students in South Korea: a descriptive study. *Child Health Nurs Res* 2023;29(4):290-9. [CrossRef]
18. Tao W, Yang J, Qu X. Utilization of, perceptions on, and intention to use ai chatbots among medical students in china: national cross-sectional study. *JMIR Med Educ* 2024;10:e57132. [CrossRef]
19. Almaiah MA, Alamri MM, Al-Rahmi W. Applying the UTAUT Model to explain the students' acceptance of mobile learning system in higher education. *IEEE Access* 2019;7:174673-86. [CrossRef]
20. García Botero G, Nguyet DA, García Botero J, Zhu C, Questier F. Acceptance and use of mobile-assisted language learning by higher education language teachers. *Lenguaje*. 2022;50(1):66-92. [CrossRef]
21. Tabachnick BG, Fidell LS. *Using Multivariate Statistics* Title: Using Multivariate Statistics <https://lccn.loc.gov/2017040173> Accessed 29 Apr, 2019.
22. Labrague LJ, Sabei S Al. Integration of AI-powered chatbots in nursing education: A scoping review of their utilization, outcomes, and challenges. *Teaching and Learning in Nursing* 2024;20(1):285-93. [CrossRef]
23. Kochis M, Parsons MQ, Rothman D, Petrusa E, Phitayakorn R. Medical students' perceptions and use of educational technologies and artificial intelligence chatbots as educational resources. *Global Surg Educ* 2024;3(1):94. [CrossRef]
24. Allison J, Hwang GJ, Mayer RE, Pellas N, Karnalim O, de Freitas S, et al. From generative AI to extended reality: multidisciplinary perspectives on the challenges, opportunities, and future of educational computing. *Journal of Educational Computing Research* 2025;63:1327-63. [CrossRef]
25. Almarwani AM. Adoption of AI in nursing education- A systematic review of factors influencing student intentions. *Appl Nurs Res* 2026;88:152068. [CrossRef]
26. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: Toward a unified view. *MIS Q* 2003;27:425-78. [CrossRef]
27. Hailu DT, Melaku MS, Abebe SA, Walle AD, Tilahun KN, Gashu KD. A modified UTAUT model for acceptance to use telemedicine services and its predictors among healthcare professionals at public hospitals in North Shewa Zone of Oromia Regional State, Ethiopia. *Front Digit Health* 2025;7:1469365. [CrossRef]



# Accuracy of Artificial Intelligence-Generated References in Dental Trauma Management

 **Esra Yıldırım Manav<sup>1</sup>**,  **Merve Özdemir<sup>2</sup>**

<sup>1</sup>Department of Restorative Dentistry, Faculty of Dentistry, Lokman Hekim University, Ankara, Türkiye

<sup>2</sup>Department of Pediatric Dentistry, Faculty of Dentistry, Lokman Hekim University, Ankara, Türkiye

## Abstract

**Introduction:** The objective of the study is to evaluate the accuracy of scientific references generated by artificial intelligence (AI) chatbots in response to clinical scenarios related to traumatic dental injuries (TDIs) and to determine the potential impact of reference errors on clinical decision-making.

**Methods:** This cross-sectional observational study analyzed 400 references generated by four AI chatbots (ChatGPT, Perplexity AI, Gemini, DeepSeek) in response to ten clinical prompts representing internationally recognized TDI categories. Each chatbot was instructed to retrieve recent PubMed-indexed studies and provide full bibliographic data. Reference authenticity and accuracy were verified using PubMed, Scopus, and Google Scholar. Hallucination severity was quantified using the reference hallucination score (RHS) scale (0–11). Non-parametric statistics and generalized linear modeling were applied ( $\alpha=0.05$ ).

**Results:** Significant differences in RHS were observed between chatbots ( $p<0.001$ ). ChatGPT and Perplexity AI demonstrated significantly lower hallucination severity compared with Gemini and DeepSeek ( $p<0.001$ ). Trauma category showed no significant effect on RHS ( $p>0.05$ ). Internal consistency for RHS components was acceptable to excellent (Cronbach's  $\alpha=0.82$ ).

**Discussion and Conclusion:** Although AI chatbots may provide rapid guidance for TDI management, the reliability of their generated references varies considerably across models. The presence of fabricated or inaccurate citations represents a potential risk for evidence-based clinical decision-making.

**Keywords:** Artificial intelligence; Bibliographic accuracy; Chatbot; Dental trauma; Evidence-based dentistry; Reference hallucination

**T**raumatic dental injuries (TDIs) represent one of the most challenging emergencies in dental practice, often requiring rapid diagnosis and evidence-based intervention to preserve pulp vitality, periodontal support, tooth function, and esthetics. The prognosis of traumatized teeth is strongly influenced by the accuracy of initial assessment, correct classification of

the injury, and adherence to current recommendations such as those issued by the International Association of Dental Traumatology (IADT).<sup>[1–3]</sup> However, because TDIs encompass a wide spectrum of clinical presentations – including avulsion, luxation, and root fractures – uncertainty regarding optimal management is common, even among experienced clinicians.<sup>[4]</sup>

**Cite this article as:** Yıldırım Manav E, Özdemir M. Accuracy of Artificial Intelligence-Generated References in Dental Trauma Management. Lokman Hekim Health Sci 2026;6(2):221–227.

**Correspondence:** Esra Yıldırım Manav, M.D. Lokman Hekim Üniversitesi, Diş Hekimliği Fakültesi, Restoratif Diş Hekimliği Anabilim Dalı, Ankara, Türkiye  
**E-mail:** esra.manav@lokmanhekim.edu.tr **Submitted:** 04.02.2026 **Revised:** 24.02.2026 **Accepted:** 03.04.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



In parallel with advances in digital health, artificial intelligence (AI) chatbots based on large language models (LLMs) are increasingly being used by clinicians and students as accessible sources of clinical information. Systems such as ChatGPT (OpenAI, USA), Perplexity AI (Perplexity Inc., USA), Gemini (Google DeepMind, USA), and DeepSeek (DeepSeek AI, China) are capable of generating structured responses to diagnostic and treatment-related queries and frequently provide supporting scientific references, thereby mimicking expert consultation.<sup>[5-7]</sup> While these tools have the potential to enhance clinical decision-making and facilitate rapid access to knowledge, their bibliographic reliability remains uncertain – particularly in acute conditions such as dental trauma, where treatment outcomes are time-sensitive.

A major concern is the phenomenon of reference hallucination, whereby AI models generate fabricated or partially incorrect citations that appear scientifically credible but do not correspond to verifiable publications.<sup>[8,9]</sup> Such inaccuracies may involve nonexistent article titles, incorrect author names, or invalid journal and digital object identifier (DOI) information. In medicine and dentistry, hallucinated references may mislead clinicians, distort evidence interpretation, and undermine academic integrity.<sup>[10-12]</sup> Reported hallucination rates exceeding 50–60% in general biomedical settings suggest that this issue may be widespread.<sup>[13]</sup> However, to date, there is no structured assessment of AI-generated reference accuracy specifically within dental traumatology – a discipline where erroneous evidence may directly affect emergency management.

Bibliographic accuracy may also vary across AI systems depending on their architecture. Retrieval-augmented models, such as ChatGPT-5 and Perplexity AI, incorporate live search or database grounding, which may reduce hallucination risk. In contrast, closed-source generative systems such as DeepSeek rely primarily on probabilistic language prediction and may therefore produce fabricated references more frequently.<sup>[14,15]</sup> Whether these architectural differences translate into measurable disparities in citation reliability during trauma-related clinical querying remains unknown.

To address this evidence gap, the primary aim of the present study was to evaluate the accuracy and reliability of AI-generated references for standardized TDI scenarios using the reference hallucination score (RHS) framework.<sup>[1,2]</sup> A secondary objective was to compare four widely used AI chatbots in terms of the frequency and severity of reference hallucinations and to determine whether

hallucination patterns differ according to chatbot model or trauma category. The null hypothesis was that no statistically significant differences would exist in RHS values among the evaluated chatbots, irrespective of trauma type or system architecture.

## Materials and Methods

### Study Design

This cross-sectional observational study was conducted in December 2025 to evaluate the accuracy of bibliographic references generated by AI chatbots in response to standardized dental trauma management scenarios. Ethics committee approval was not required because the study did not include human subjects, patient information, or biological materials. The methodological framework was based on the validated RHS system proposed by Aljamaan et al.<sup>[13]</sup> and later adapted for dental research reliability assessment.<sup>[2]</sup>

### Selection of AI Chatbots

Four publicly accessible, English-language AI chatbots were selected based on their global academic usage, technological diversity, and public availability. The evaluated systems included ChatGPT (version 5.2), Perplexity AI, Gemini (3 Flash), and DeepSeek, all of which were accessed in their most recent publicly available versions at the time of data collection.

Each chatbot was accessed through its official web interface using a verified account on a secure institutional network. All sessions were conducted in incognito mode to prevent previous chat memory from influencing outputs.

### Sample Size and Power Calculation

An a priori power analysis was performed using GPower 3.1.9.6\* (Heinrich-Heine University, Düsseldorf, Germany). Assuming a medium effect size ( $f=0.25$ ), significance level  $\alpha=0.05$ , and statistical power  $(1-\beta)=0.80$ , the minimum required total sample size was estimated at 160 references ( $\approx 40$  per chatbot). The assumption of a medium effect size was based on conventional benchmarks proposed by Cohen and was consistent with a previous study evaluating the performance of LLM-based chatbots in dentistry.<sup>[16]</sup> Given the absence of established benchmarking data specific to AI-based guideline adherence in dental traumatology, a medium effect size was considered a methodologically appropriate and conservative estimate to detect practically meaningful differences between chatbot models. Since the present study analyzed 400 references (4 chatbots  $\times$  10 prompts  $\times$  10

references), the achieved power exceeded 0.95, confirming adequate sample size for reliable inter-model comparisons.

### Prompt Development

Ten standardized clinical prompts were developed to represent the most common types of TDIs as classified by the IADT guidelines.<sup>[3-5]</sup> Each prompt simulated a realistic case scenario including patient age, injury type, and clinical question. The prompts requested both a short management summary and a list of ten supporting literature references in Vancouver style.

Each scenario prompt consisted of (i) a clinical vignette and (ii) a standardized instruction block applied identically across all chatbots. This block required the model to (a) begin by searching PubMed; (b) select 10 recent and relevant articles; and (c) report, for each article, the title, authors, journal, publication date, citation count, DOI, web link, and PubMed link, formatted consistently. The full prompt templates are provided in Appendix 1.

Each chatbot received the same ten prompts in identical order and formatting. All interactions were carried out by a single experienced investigator (EYM) to ensure standardization and consistency. All responses were exported as plain text and anonymized for evaluation.

### Reference Verification

All references generated by the chatbots were manually verified using PubMed, Scopus, and Google Scholar. Each reference was checked for existence, bibliographic accuracy, and relevance to the prompted topic. If a reference could not be found in any of the databases or exhibited falsified information (e.g., fabricated title, incorrect author list, or non-existent DOI), it was classified as hallucinated. Minor discrepancies, such as incorrect publication year or typographical errors, were considered partial hallucinations.

### Scoring Criteria

The RHS was applied to quantify hallucination severity for each citation across seven bibliographic identifiers:

1. Title
2. Authors' names
3. Journal name
4. Publication year
5. Digital object identifier (DOI)
6. Web link (URL)
7. Relevance to the trauma topic.

Each major hallucination (e.g., incorrect or missing title, author list, journal, or DOI) received 2 points, while minor hallucinations (e.g., wrong year, invalid link, or irrelevant topic) received 1 point. Thus, the total RHS per reference ranged from 0 (fully accurate) to 11 (completely hallucinated). All references were independently evaluated by two calibrated reviewers with expertise in dental traumatology. The reviewers were blinded to the chatbot identity during scoring. Each citation was assessed according to the predefined RHS criteria. In cases of disagreement, the reference was re-evaluated through discussion, and consensus was reached. Inter-rater reliability for RHS scoring was assessed using the intraclass correlation coefficient (ICC), demonstrating excellent agreement (ICC=0.93).

A mean RHS value was calculated for each chatbot and for each trauma category.

Higher scores indicated greater factual inaccuracy. All references were assessed independently of whether the corresponding textual answer was clinically accurate or inaccurate. Thus, hallucination scoring reflected citation validity rather than clinical reasoning quality.

### Statistical Analysis

Statistical analyses were performed using IBM Statistical Package for the Social Sciences Statistics v29 (IBM Corp., Armonk, NY, USA). Data normality was assessed using the Kolmogorov–Smirnov test, and continuous variables were expressed as mean±standard deviation. Inter-model differences in RHS values were examined using the Kruskal–Wallis test, followed by Bonferroni-adjusted pairwise post hoc comparisons where appropriate. Differences in RHS values across trauma categories were analyzed using the Mann–Whitney U test or Kruskal–Wallis test, depending on the number of groups being compared. To identify independent predictors of hallucination severity, a Generalized Linear Model (gamma distribution, log-link function) was constructed, with chatbot type and trauma category entered as fixed factors. Internal consistency across RHS components was evaluated using Cronbach's  $\alpha$  coefficient. A  $p<0.05$  was considered statistically significant.

### Results

Table 1 shows that there were marked differences in RHS scores between the evaluated chatbots. A statistically significant difference in RHS values among the four chatbots was confirmed using the Kruskal–Wallis test ( $p<0.001$ ).

**Table 1.** Reference the hallucination score across the evaluated artificial intelligence-based chatbot models

Chatbot	Mean±SD	Median (IQR)	Min–Max
ChatGPT	3.92±2.63	4.0 (2.0–6.0)	0–10
Perplexity AI	4.65±3.04	4.0 (2.75–6.0)	0–10
Gemini	6.21±3.62	6.5 (4.0–10.0)	0–10
DeepSeek	7.18±3.79	10.0 (5.0–10.0)	0–10
p	<0.001		

AI: Artificial intelligence; SD: Standard deviation; IQR: Interquartile range; Min: Minimum; Max: Maximum. Kruskal–Wallis test:  $\chi^2=51.9$

Pairwise comparisons showed that both ChatGPT and Perplexity AI produced significantly lower RHS values than Gemini 3 Flash and DeepSeek (ChatGPT vs. Gemini:  $p<0.001$ ; ChatGPT vs. DeepSeek:  $p<0.001$ ; Perplexity AI vs. DeepSeek:  $p<0.001$ ; Perplexity AI vs. Gemini:  $p=0.010$ ). However, only a borderline difference was observed between ChatGPT and Perplexity AI ( $p=0.05$ ), and Gemini and DeepSeek also did not differ significantly from one another ( $p>0.05$ ).

RHS scores were also examined according to trauma category, summarized in Table 2. Although small numerical variations were observed, interquartile ranges overlapped considerably, and there were no statistically significant differences in RHS values across trauma categories ( $p>0.05$ ) (Fig. 1).

To further assess predictors of hallucination severity, a generalized linear model with gamma distribution and log-link function was constructed including chatbot type and trauma category as fixed effects. As summarized in Table 3, chatbot type remained an independent predictor of RHS, with DeepSeek and Gemini demonstrating significantly higher hallucination severity than ChatGPT ( $p<0.001$ ).

Internal consistency analysis of the seven RHS components demonstrated acceptable to excellent reliability. As shown in Table 4, Cronbach's  $\alpha$  was 0.82 for the total dataset. Model-specific  $\alpha$  values were 0.68 for ChatGPT, 0.76 for Perplexity AI, 0.82 for Gemini, and 0.88 for DeepSeek.

**Table 2.** Reference the hallucination score across dental trauma categories

Trauma category	Mean±SD	Median (IQR)
Avulsion	6.28±3.93	7.0 (2.0–10.0)
Intrusion	5.87±3.66	6.0 (4.0–8.5)
Extrusion	5.78±3.51	6.0 (4.0–8.5)
Lateral luxation	5.41±3.47	6.0 (3.0–8.0)
Subluxation	5.00±3.64	5.0 (3.0–7.0)
Uncomplicated crown fracture	5.33±3.29	6.0 (3.0–7.0)
Complicated crown fracture	5.54±3.45	6.0 (3.0–8.0)
Root fracture	5.72±3.68	6.0 (3.0–8.5)
Alveolar fracture	5.83±3.74	6.0 (3.0–9.0)
Post-traumatic pulp necrosis	5.69±3.59	6.0 (3.0–8.0)
p	0.876	

SD: Standard deviation; IQR: Interquartile range; Kruskal–Wallis test:  $\chi^2=4.49$

## Discussion

This study evaluated the bibliographic reliability of AI-based chatbots when generating scientific references for standardized dental trauma scenarios. Significant inter-model differences in RHS were observed, whereas trauma category did not significantly influence hallucination severity. The null hypothesis was therefore partially rejected. While statistically significant differences were identified between ChatGPT and other systems – particularly Gemini and DeepSeek – no significant differences were observed among Perplexity AI, Gemini, and DeepSeek. These findings indicate that hallucination severity varies across individual LLMs, although such differences are not uniformly distributed across all chatbot architectures.

The generalized linear model further confirmed chatbot type as an independent predictor of RHS values, reinforcing the robustness of the inter-model comparison. In contrast, the trauma category did not emerge as a significant factor, suggesting that reference instability reflects model-specific characteristics rather than clinical scenario complexity. Whether the prompt concerned avulsion, luxation, fracture, or post-traumatic pulp necrosis, hallucination severity remained relatively stable.

**Table 3.** Generalized linear model (gamma distribution, log-link) assessing predictors of reference hallucination score

Predictor comparison	$\beta$ (SE)	Exp( $\beta$ )	95% CI for Exp( $\beta$ )	z	p
Perplexity AI versus ChatGPT	0.171 (0.087)	1.19	1.00–1.41	1.96	0.050
Gemini versus ChatGPT	0.463 (0.087)	1.59	1.34–1.89	5.31	<0.001
DeepSeek versus ChatGPT	0.608 (0.087)	1.84	1.55–2.18	6.97	<0.001

$\beta$ =regression coefficient; SE: Standard error; Exp( $\beta$ ): Rate ratio; CI: Confidence interval. Reference: ChatGPT-5.2.

**Table 4.** Internal consistency of reference hallucination score components across chatbots

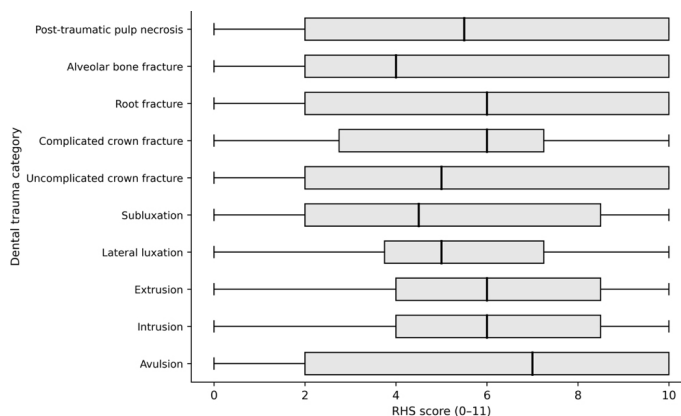
Chatbot	Cronbach's $\alpha$
ChatGPT	0.68
Perplexity AI	0.76
Gemini	0.82
DeepSeek	0.88
Overall	0.82

Cronbach's  $\alpha$  was used to assess the internal consistency of reference hallucination score components across chatbots.  $\geq 0.90$ : excellent, 0.80–0.89: good, 0.70–0.79: acceptable,  $< 0.70$ : questionable/poor

These findings support emerging evidence that reference hallucination is a model-dependent phenomenon, primarily shaped by system architecture and retrieval mechanisms rather than by content domain.<sup>[8,9,13–17]</sup> Consistent with prior biomedical AI research, retrieval-augmented systems generated references with fewer inaccuracies than models relying predominantly on generative language prediction. In the present analysis, ChatGPT and Perplexity AI demonstrated significantly lower RHS values compared with Gemini and DeepSeek, suggesting comparatively stronger grounding in indexed databases. These observations align with previous reports indicating that integration of external search functions reduces the likelihood of fabricated bibliographic content.<sup>[6,10–12,18–20]</sup> Conversely, generative-only models continued to produce plausible yet unverifiable citations, reinforcing concerns raised in earlier medical and dental AI literature.<sup>[10–12,21–24]</sup>

The study also demonstrated acceptable to good internal consistency of the RHS framework across chatbot outputs, supporting its utility as a structured metric for benchmarking citation reliability. These findings are consistent with the original validation of RHS in medical AI settings.<sup>[13]</sup> The reproducibility of scoring across systems strengthens the methodological validity of the present comparisons.

Importantly, the present investigation focused exclusively on bibliographic accuracy rather than the clinical correctness of chatbot-generated management recommendations. Therefore, the findings should be interpreted as evidence of variability in citation reliability rather than as an assessment of clinical reasoning quality. In time-sensitive disciplines such as dental traumatology, unreliable or unverifiable citations may complicate rapid evidence verification processes, underscoring the importance of independent source validation. However, the study did not evaluate guideline concordance or treatment accuracy, and such dimensions require dedicated clinical evaluation frameworks.<sup>[25–27]</sup>

**Figure 1.** Boxplot distribution of reference hallucination scores across dental trauma categories. The median, interquartile range, and minimum–maximum values are shown for each trauma type.

From a broader academic perspective, the results contribute to ongoing discussions regarding the responsible integration of AI technologies into healthcare research and education. Although retrieval-augmented systems performed comparatively better, none achieved complete bibliographic accuracy. This indicates that manual verification remains essential when AI-generated references are used for academic writing or clinical support. Established guideline documents, such as those issued by the IADT,<sup>[1–3]</sup> continue to represent the authoritative standard for trauma management.

The present study has several strengths, including standardized trauma-based prompts aligned with internationally recognized classifications, a substantial reference sample, and multi-database verification. Nevertheless, limitations should be acknowledged. First, only English-language, general-purpose chatbots were evaluated at a single time point; ongoing model updates may influence performance. Second, the analysis was confined to bibliographic verification and did not examine the clinical validity of AI-generated recommendations. Third, findings cannot be extrapolated to specialized AI systems trained on curated medical datasets.

Future research should incorporate longitudinal designs to assess the stability of reference accuracy across model updates. Comparative studies involving domain-specific or medically trained AI platforms may clarify whether curated training data reduce hallucination rates. In addition, dedicated investigations into clinical guideline concordance are warranted to determine whether bibliographic hallucination correlates with clinical inaccuracy. Expanding analyses to multilingual settings and real-world clinical prompts would further enhance generalizability and practical relevance.

## Conclusion

AI-based chatbots showed marked variability in the accuracy of the references they generated for dental trauma scenarios. Retrieval-augmented systems demonstrated comparatively lower hallucination severity; however, none of the evaluated models achieved full bibliographic reliability. Reference hallucination therefore appears to be a model-dependent rather than a context-dependent phenomenon. Because evidence-based guidance is essential for the prognosis of TDIs, inaccurate or unverifiable AI-generated citations may pose challenges for evidence verification processes. However, the present findings do not extend to clinical decision-making accuracy, which requires dedicated investigation. AI chatbots should therefore be regarded as supportive tools rather than independent bibliographic resources, and all AI-generated references should be independently verified before clinical or academic use.

**Ethics Committee Approval:** Ethics committee approval was not required because the study did not include human subjects, patient information, or biological materials.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** During the preparation of this manuscript, the authors used artificial intelligence-based language models solely to improve language clarity, grammar, and overall readability. The AI tools were not used for data analysis, data interpretation, or the generation of scientific content. All conceptualization, study design, data collection, statistical analysis, and interpretation of findings were performed exclusively by the authors. The authors critically reviewed, edited, and take full responsibility for the final content of the manuscript.

**Authorship Contributions:** Concept: EYM; Design: EYM, MO; Supervision: MO; Materials: EYM, MO; Data collection and/or processing: EYM; Analysis and/or interpretation: EYM; Literature review: EYM; Writing: EYM, MO; Critical review: EYM, MO.

**Peer-review:** Double blind peer-reviewed.

## References

1. Fouad AF, Abbott PV, Tsilingaridis G, Cohenca N, Lauridsen E, Bourguignon C, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 2. Avulsion of permanent teeth. *Dent Traumatol* 2020;36(4):331-42. [\[CrossRef\]](#)
2. Day PF, Flores MT, O'Connell AC, Abbott PV, Tsilingaridis G, Fouad AF, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 3. Injuries in the primary dentition. *Dent Traumatol* 2020;36(4):343-59. [\[CrossRef\]](#)
3. Levin L, Day PF, Hicks L, O'Connell A, Fouad AF, Bourguignon C, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: General introduction. *Dent Traumatol* 2020;36(4):309-13. [\[CrossRef\]](#)
4. Magno MB, Nadelman P, Leite KLDF, Ferreira DM, Pithon MM, Maia LC. Associations and risk factors for dental trauma: a systematic review of systematic reviews. *Community Dent Oral Epidemiol* 2020;48(6):447-63. [\[CrossRef\]](#)
5. Gao S, Wang X, Xia Z, Zhang H, Yu J, Yang F. Artificial intelligence in dentistry: a narrative review of diagnostic and therapeutic applications. *Med Sci Monit* 2025;31:e946676. [\[CrossRef\]](#)
6. Liu TY, Lee KH, Mukundan A, Karmakar R, Dhiman H, Wang HC. AI in dentistry: innovations, ethical considerations, and integration barriers. *Bioengineering (Basel)* 2025;12(9):928. [\[CrossRef\]](#)
7. Ghaffari M, Zhu Y, Shrestha A. A review of advancements of artificial intelligence in dentistry. *Dent Rev (Heidelb)* 2024;4(2):100081. [\[CrossRef\]](#)
8. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv* 2023;55:1-38. [\[CrossRef\]](#)
9. Rawte V, Sheth A, Das A. A survey of hallucination in large language models. *arXiv* 2023.
10. Munaye YY, Admass W, Belayneh Y, Molla A, Asmare M. ChatGPT in education: a systematic review on opportunities, challenges, and future directions. *Algorithms* 2025;18(6):352. [\[CrossRef\]](#)
11. Kotsis KT. Scientific authorship in the age of AI: challenges for editors and institutions. *Eur J Manag Econ Bus* 2025;2(6):209-16. [\[CrossRef\]](#)
12. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LTJ, Li YCJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J Biomed Sci* 2025;32(1):45. [\[CrossRef\]](#)
13. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Med Inform* 2024;12(1):e54345. [\[CrossRef\]](#)
14. Gorelik AJ, Li M, Hahne J, Wang J, Ren Y, Yang L, et al. Ethics of AI in healthcare: a scoping review demonstrating applicability of a foundational framework. *Front Digit Health* 2025;7:1662642. [\[CrossRef\]](#)
15. Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. *JAMA Ophthalmol* 2023;141:819-24. [\[CrossRef\]](#)
16. Kandemir M, Saribaş EE. Comparative performance of large language models in answering periodontology questions from the Turkish Dental Specialty Examination: a cross-sectional study on accuracy and coverage. *BMC Oral Health* 2025;25:1804. [\[CrossRef\]](#)
17. Künzle P, Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. *Clin Oral Investig* 2024;28:575. [\[CrossRef\]](#)

18. Turan Gökdoğan C, Arılı Öztürk E, Aktaş Ş, Çanakçı BC. Comparison of chatbots' accuracy in endodontics questions in dentistry specialization exam in Türkiye: ChatGPT-4o, Gemini Advanced, Copilot, and Claude. *BMC Oral Health* 2025;26(1):28. [\[CrossRef\]](#)
19. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health* 2023;5:e179-81. [\[CrossRef\]](#)
20. Zybaczynska J, Norris M, Modi S, Brennan J, Jhaveri P, Craig TJ, et al. Artificial intelligence-generated scientific literature: a critical appraisal. *J Allergy Clin Immunol Pract* 2024;12:106-10. [\[CrossRef\]](#)
21. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health* 2023;2:e0000198. [\[CrossRef\]](#)
22. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords for radiologists. *Radiology* 2023;307:e230163. [\[CrossRef\]](#)
23. Alyasiri OM, Salman AM, Akhtom DA, Salisu S. ChatGPT revisited: using ChatGPT-4 for finding references and editing language in medical scientific articles. *J Stomatol Oral Maxillofac Surg* 2024;125:101842. [\[CrossRef\]](#)
24. Goktas P, Grzybowski A. Assessing the impact of ChatGPT in dermatology: a comprehensive rapid review. *J Clin Med* 2024;13:5909. [\[CrossRef\]](#)
25. Pham T. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. *R Soc Open Sci* 2025;12(5):241873. [\[CrossRef\]](#)
26. Küçük Keleş Ö, Arslan ZB. Performance of artificial intelligence chatbots in the diagnosis and management of simulated dental trauma cases: an evaluation based on IADT guidelines. *Clin Oral Investig* 2026;30:26. [\[CrossRef\]](#)
27. Gonzalez-Valenzuela RE, Mettes P, Loos BG, Marquering H, Berkhout E. Accuracy of deep learning-based AI models for early caries lesion detection: the influence of annotation quality and reference choice. *Clin Oral Investig* 2025;29:598. [\[CrossRef\]](#)

## **Appendix 1**

### **Prompt 1 :**

A 10-year-old patient presents with an avulsed permanent maxillary incisor. According to current evidence-based guidelines, describe the recommended immediate management, splinting protocol, endodontic considerations, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

### **Prompt 2:**

A 9-year-old patient presents with an intruded permanent maxillary central incisor. According to current evidence-based guidelines, describe the recommended management approach, follow-up protocol, possible complications, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

### **Prompt 3:**

A patient presents with an extruded permanent tooth following dental trauma. According to current evidence-based guidelines, describe the emergency management, repositioning technique, splinting protocol, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).

3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 4:

A patient presents with a laterally luxated permanent tooth with alveolar socket displacement. According to current evidence-based guidelines, describe the recommended repositioning, stabilization, endodontic considerations, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 5:

A patient presents with subluxation of a permanent tooth following trauma. According to current evidence-based guidelines, describe the recommended clinical management, monitoring protocol, possible complications, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.

2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 6:

A patient presents with an uncomplicated enamel–dentin crown fracture without pulp exposure. According to current evidence-based guidelines, describe the recommended treatment approach, restorative options, follow-up protocol, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 7:

A patient presents with a complicated crown fracture with pulp exposure. According to current evidence-based guidelines, describe the recommended vital pulp therapy or endodontic treatment approach, restorative procedures, follow-up, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 8:

A patient presents with a horizontal root fracture of a permanent tooth. According to current evidence-based

guidelines, describe the diagnostic procedures, repositioning and splinting protocol, endodontic considerations, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 9:

A patient presents with an alveolar bone fracture associated with dental trauma. According to current evidence-based guidelines, describe the stabilization procedure, splinting duration, follow-up protocol, possible complications, and prognosis. Please provide 10 supporting references.

Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.
6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.

Prompt 10:

A patient develops post-traumatic pulp necrosis following dental trauma. According to current evidence-based guidelines, describe the indications for endodontic intervention, treatment protocol, follow-up recommendations, and prognosis. Please provide 10 supporting references.

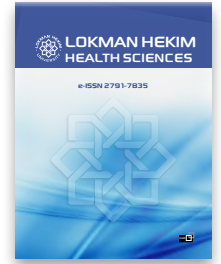
Include the following information for each article:

1. Article title.
2. Author(s).
3. Journal name.
4. Date of publication.
5. Number of citations.

6. DOI.
7. Web link to the article.
8. PubMed link.

Instructions:

1. Begin by searching PubMed.
2. Review the search results and select ten recent articles that are relevant.
3. Ensure that all information is accurate and up to date.
4. Format the list of articles in a clear and organized manner, using a consistent style for each entry.
5. Include any additional information or notes that may be relevant or helpful for readers.
6. Double-check the accuracy and completeness of the list before publishing or submitting it.



## ORIGINAL ARTICLE

# Bibliometric Trends of Artificial Intelligence Applications in Pediatric Dentistry: A Scopus and Web of Science Analysis (2000–2024)

 Hüllya Çerçi Akçay

Department of Pediatric Dentistry, Kocaeli Health and Technology University, Kocaeli, Türkiye

## Abstract

**Introduction:** Artificial intelligence (AI) has gained increasing attention in dentistry, particularly in diagnostic imaging, caries detection, and orthodontic prediction. However, pediatric applications remain underexplored, and no bibliometric synthesis has specifically mapped this research field.

**Methods:** A bibliometric analysis was performed using Scopus and Web of Science Core Collection (2000–2024). Search terms included “artificial intelligence,” “machine learning,” “deep learning,” and “pediatric dentistry.” Bibliographic data were analyzed using VOSviewer (v.1.6.20) and Microsoft Excel, focusing on co-authorship networks (authors and countries), keyword co-occurrence networks, citation and co-citation analyses, and bibliographic coupling of publications.

**Results:** A total of 92 publications were identified (Scopus: 53; WoS: 39). Annual outputs increased sharply after 2020, peaking in 2024. India, Turkey, and the United States were the most productive countries, while collaborations across regions remained limited. Keyword clustering revealed three thematic areas: (i) AI-driven diagnostics, (ii) caries risk prediction, and (iii) orthodontic growth assessment. Preventive and behavioral domains were notably underrepresented.

**Discussion and Conclusion:** Research on AI in pediatric dentistry is rapidly growing but remains diagnostically focused. This study provides the first pediatric dentistry-focused bibliometric overview of AI research, explicitly mapping publication trends, thematic concentrations, and underexplored preventive and behavioral domains. Future research should extend beyond diagnostics, strengthen international collaboration, and address ethical considerations to support responsible clinical integration.

**Keywords:** Artificial intelligence; Behavior management; Bibliometric analysis; Diagnostic imaging; Pediatric dentistry

Artificial intelligence (AI) has rapidly transformed the landscape of healthcare research and clinical practice, offering novel opportunities for diagnostics, decision support, and personalized treatment planning. In dentistry, AI-based technologies such as machine learning and deep

learning have been increasingly employed for diagnostic imaging, caries detection, treatment planning, and risk prediction.<sup>[1–3]</sup> These advancements hold particular promise in pediatric dentistry, where early diagnosis and intervention are critical to long-term oral health outcomes.<sup>[4,5]</sup>

**Cite this article as:** Çerçi Akçay H. Bibliometric Trends of Artificial Intelligence Applications in Pediatric Dentistry: A Scopus and Web of Science Analysis (2000–2024). Lokman Hekim Health Sci 2026;6(2):228–236.

**Correspondence:** Hüllya Çerçi Akçay, M.D. Kocaeli Sağlık ve Teknoloji Üniversitesi, Çocuk Diş Hekimliği Anabilim Dalı, Kocaeli, Türkiye

**E-mail:** hulya.akcay@kocaelisaglik.edu.tr **Submitted:** 07.09.2025 **Revised:** 24.04.2026 **Accepted:** 27.04.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



Within pediatric dentistry, the integration of AI has gained momentum in domains such as caries risk assessment, orthodontic growth prediction, pulp therapy decision-making, and behavior management strategies.<sup>[6,7]</sup> For example, deep learning algorithms have demonstrated superior accuracy in detecting dental caries and anomalies in panoramic radiographs, often surpassing conventional diagnostic approaches.<sup>[8]</sup> Similarly, machine learning models have been applied to predict treatment outcomes in interceptive orthodontics and evaluate craniofacial growth, offering pediatric dentists evidence-based decision support tools.<sup>[9]</sup> Beyond diagnostic accuracy, these applications also hold clinical relevance for preventive strategies and behavior management in pediatric dentistry, where early risk identification and child-centered care are essential components of long-term oral health outcomes. For instance, AI-supported tools that assist in anticipating dental anxiety or tailoring communication strategies may improve patient cooperation during clinical procedures, thereby facilitating more effective and less stressful dental care for children.

Despite the rapid growth of AI-related research in dentistry, pediatric-focused applications remain relatively limited compared to other dental specialties such as orthodontics, prosthodontics, or maxillofacial radiology.<sup>[10,11]</sup>

Furthermore, the distribution of research productivity appears uneven across regions, with countries such as the United States, China, and India leading in AI innovation, while contributions from developing nations remain sparse.<sup>[12,13]</sup>

These disparities underscore the need to systematically map the global research landscape to identify key contributors, thematic foci, and emerging trends. Bibliometric analysis provides a robust methodological framework for evaluating research trends, citation impact, and collaboration patterns in a given field.<sup>[14,15]</sup> By quantifying publication output, analyzing co-authorship and country networks, and visualizing keyword co-occurrence, bibliometric studies allow researchers to understand how a topic evolves and where potential research gaps exist.<sup>[16]</sup>

Previous bibliometric studies in dentistry have mapped AI research in endodontics, implantology, and general oral health, but none have comprehensively evaluated pediatric dentistry.<sup>[7–9]</sup> The absence of a focused bibliometric synthesis makes it unclear which countries, authors, or themes are leading the field and where research gaps exist.

Therefore, this study aims to systematically map AI research in pediatric dentistry by analyzing publications indexed in Scopus and Web of Science between 2000 and 2024.

Specifically, we assess (i) Publication and citation trends, (ii) leading authors and countries, (iii) thematic clusters of research through keyword analysis, and (iv) collaboration networks.

## Materials and Methods

### Study Design and Data Sources

This study was designed as a bibliometric analysis to systematically map research trends related to AI applications in pediatric dentistry. Two major bibliographic databases, Scopus and the Web of Science Core Collection, were selected due to their extensive coverage of high-quality, peer-reviewed literature and their widespread use in bibliometric research.

### Search Strategy

A comprehensive and reproducible literature search was conducted on May 25, 2025, to identify studies related to AI applications in pediatric dentistry. The search strategy was developed using a combination of relevant keywords and Boolean operators to ensure adequate coverage of the literature.

The complete search queries used in each database were as follows:

- Scopus: (TITLE-ABS-KEY (“artificial intelligence” OR “machine learning” OR “deep learning”)) AND (TITLE-ABS-KEY (“pediatric dentistry” OR “pedodontics”)) AND PUBYEAR > 1999 AND PUBYEAR < 2025
- Web of Science Core Collection: TS=(“artificial intelligence” OR “machine learning” OR “deep learning”) AND TS=(“pediatric dentistry” OR “pedodontics”) AND PY=(2000–2024)

The search strategy incorporated Boolean operators (“AND”, “OR”) to combine AI-related and pediatric dentistry-related terms. No restrictions were applied regarding language or document type to maximize the comprehensiveness of the dataset.

### Data Extraction and Preprocessing

The initial search retrieved 92 records (Scopus: 53; Web of Science: 39). Records retrieved from both databases were merged, and duplicate entries were manually identified and removed based on comparison of titles, author names, and digital object identifiers (DOIs).

Bibliographic information, including authors, affiliations, titles, abstracts, keywords, source titles, publication years, and citation counts, was exported in CSV format. Before

analysis, data cleaning and harmonization procedures were performed to standardize author names and keywords, thereby minimizing inconsistencies that could affect network visualization and interpretation.

Duplicate records retrieved from both databases were identified and removed manually based on comparison of titles, author names, and DOIs. No automated deduplication software was used. Although the exact number of duplicate records was not recorded, the process was conducted carefully to ensure data accuracy.

#### Bibliometric Analysis

Quantitative bibliometric analyses were conducted using VOSviewer software (version 1.6.20) and Microsoft Excel. The following bibliometric indicators and networks were examined:

- Co-authorship networks (authors and countries)
- Keyword co-occurrence networks
- Citation and co-citation networks
- Bibliographic coupling of publications.

Minimum inclusion thresholds were applied to improve the interpretability of the visualizations: authors with  $\geq 2$  publications and  $\geq 20$  citations; countries with  $\geq 3$  publications and  $\geq 10$  citations; keywords with  $\geq 3$  occurrences; and journals with  $\geq 3$  publications. For each analysis, only the largest connected components were visualized to enhance clarity and reduce network fragmentation.

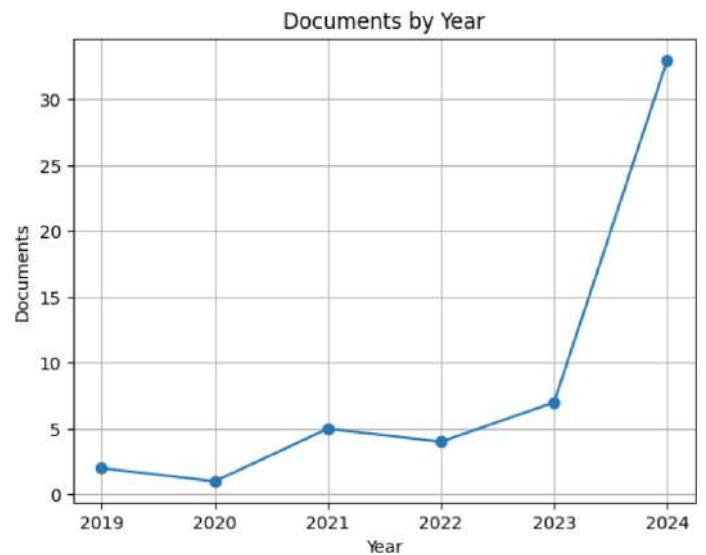
All results were generated using the full counting method, whereby each author, country, or keyword occurrence was counted equally across publications. This approach was selected to facilitate transparent comparison of productivity, collaboration patterns, and thematic structures within the bibliometric networks.

#### Ethical Considerations

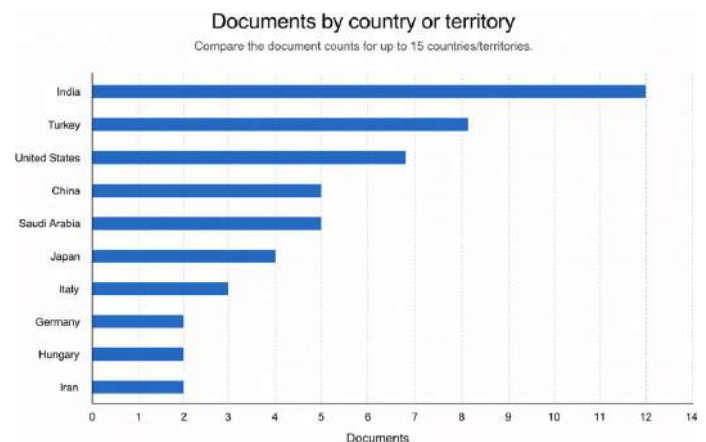
As this study relied exclusively on publicly available bibliographic metadata and did not involve human or animal participants, ethical approval was not required.

#### Results

This bibliometric analysis included a total of 92 publications (Scopus: 53; Web of Science: 39) focusing on the intersection of AI, machine learning, and pediatric dentistry between 2000 and 2024. The analysis covered publication output over time, country, and author productivity, collaboration patterns, keyword co-occurrence, and citation-based relationships.



**Figure 1.** Annual distribution of publications on artificial intelligence in pediatric dentistry based on Scopus data (2019–2024). Data were retrieved from Scopus and the figure was generated by the author.



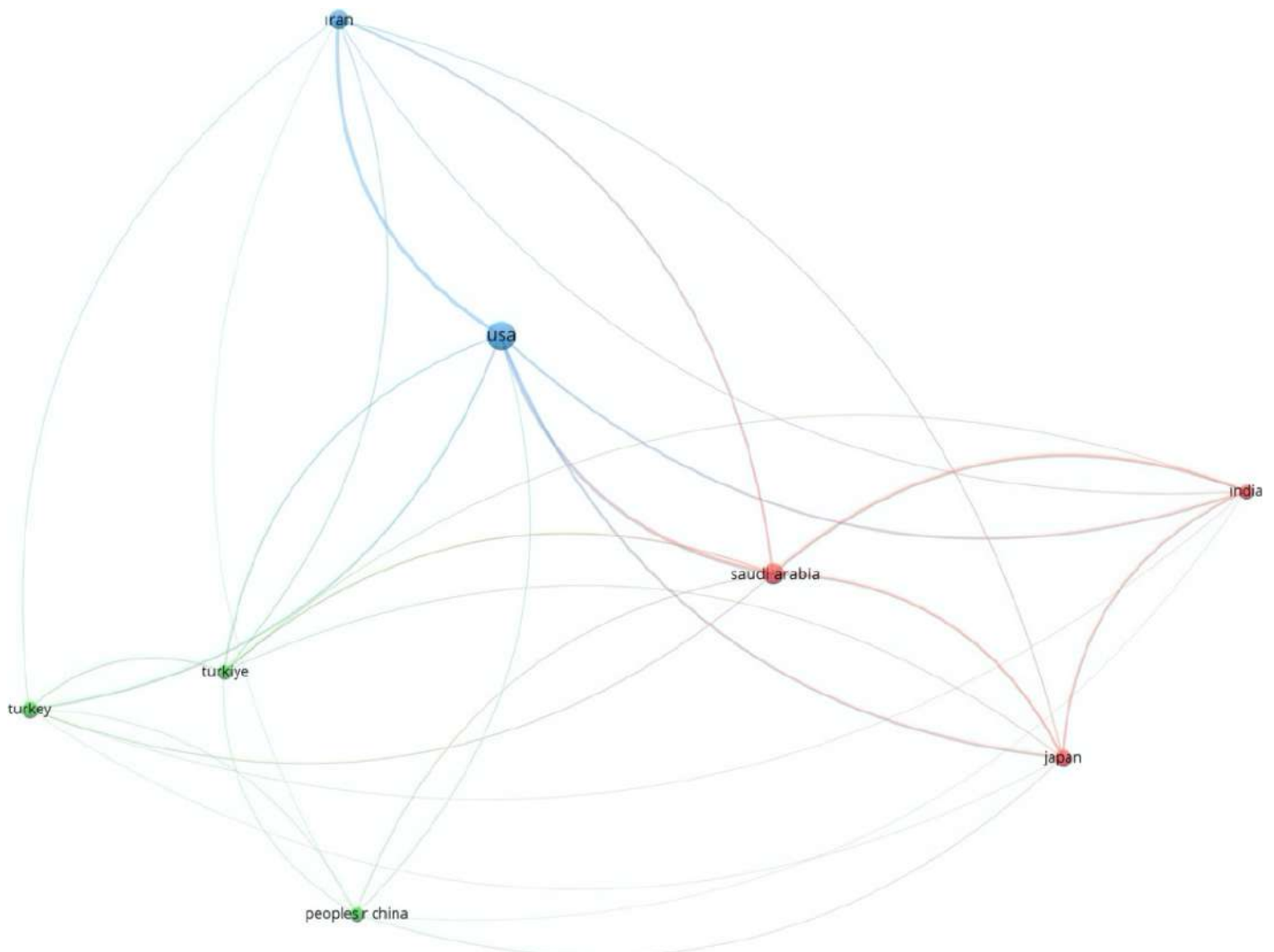
**Figure 2.** Distribution of publications on artificial intelligence in pediatric dentistry by country based on Scopus data. Data were retrieved from the Scopus database and the figure was generated by the author.

#### Publication Trends

The temporal distribution of publications showed an increase in scientific output after 2021, with a peak in 2024 (Fig. 1). Fewer than ten publications per year were identified before 2022, whereas more than thirty publications were recorded in 2024. Citation counts increased over the same period.

#### Geographical Distribution

Analysis of country-level contributions showed that India ( $n=13$ ), Turkey ( $n=9$ ), and the United States ( $n=7$ ) were the most productive countries within the Scopus dataset (Fig. 2). Additional contributions were observed from China ( $n=5$ ), Saudi Arabia ( $n=5$ ), Japan ( $n=4$ ), and several



**Figure 3.** Network visualization of international collaboration among countries in artificial intelligence research in pediatric dentistry. The visualization was generated using VOSviewer software based on bibliometric data retrieved from the Scopus database.

European countries including Italy ( $n=3$ ), Germany ( $n=2$ ), and Hungary ( $n=2$ ). Iran accounted for two publications during the study period. Data derived from the Web of Science database showed a similar distribution.

Bibliometric coupling of countries is presented in Figure 3. The mapping included multiple clusters. India, Turkey, and the United States were among the most connected countries. China, Saudi Arabia, and Japan were also represented within separate clusters.

### Authorship and Collaboration Networks

The co-authorship analysis revealed a relatively limited and fragmented collaboration structure within the field. The network consisted of a small number of interconnected clusters, indicating that research activity was concentrated among a restricted group of contributors. Overlay visualization

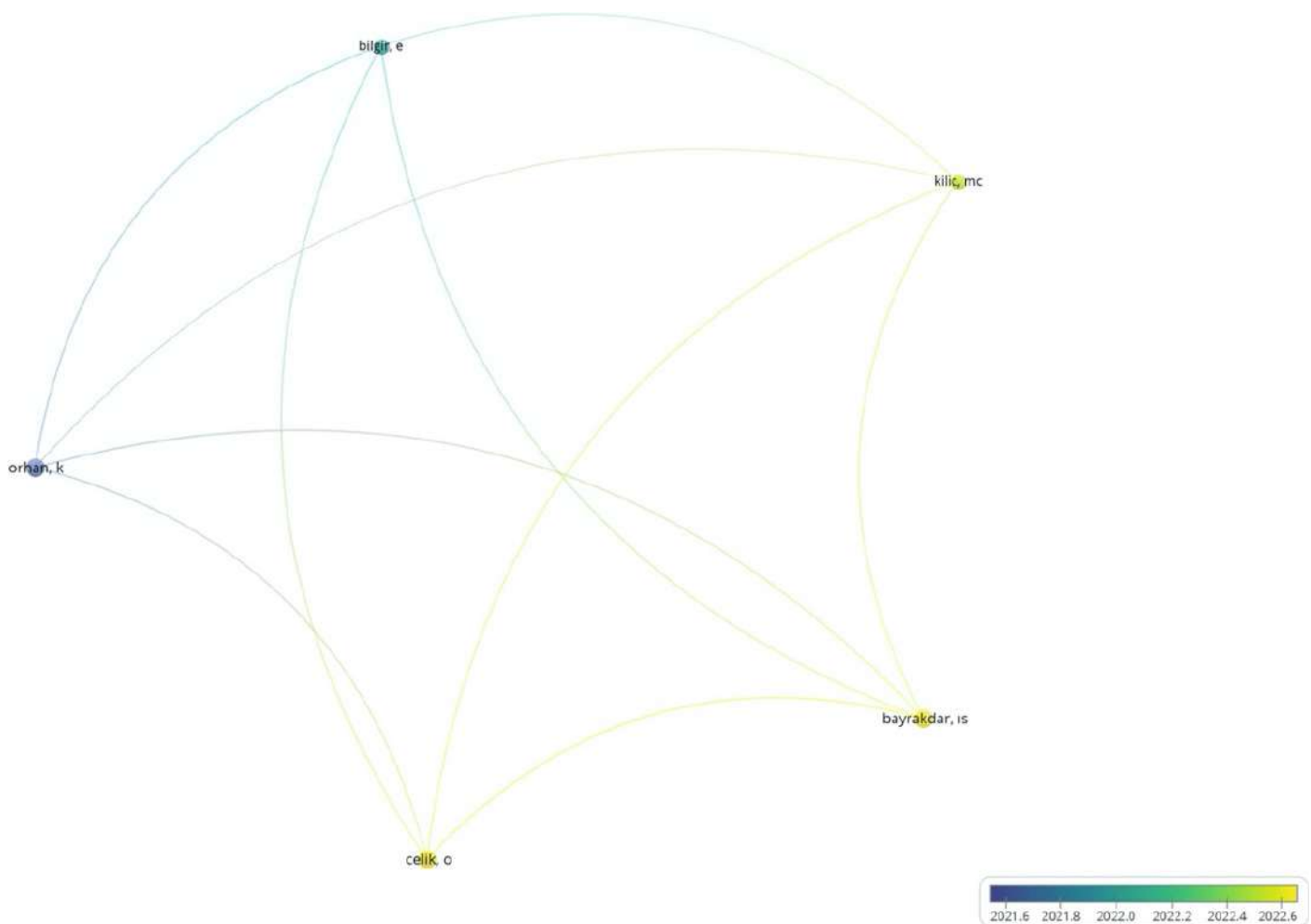
suggested that collaborative publications became more prominent during the 2021–2023 period. Overall, the network demonstrated moderate collaboration intensity with limited large-scale international connectivity (Fig. 4).

### Keyword Co-occurrence and Thematic Clusters

Keyword analysis (Fig. 5) identified “artificial intelligence,” “machine learning,” “deep learning,” “pediatric dentistry,” and “dental caries” as the most frequently occurring terms. These keywords were grouped into three clusters: (i) Diagnostic imaging and predictive modeling, (ii) Caries detection and risk assessment, and (iii) Orthodontic growth and development.

### Citation and Bibliographic Coupling Networks

Citation network analysis demonstrated the presence of



**Figure 4.** Overlay visualization of author collaboration in artificial intelligence research in pediatric dentistry. The visualization was generated using VOSviewer software based on bibliometric data retrieved from the Scopus database.

several interconnected clusters representing influential studies within the field. Larger nodes and stronger link strengths indicated that certain publications served as central reference points in the development of AI-related pediatric dentistry research. In addition, more recent studies appeared to contribute to the expansion and diversification of thematic research areas. Bibliographic coupling analysis further revealed shared intellectual foundations among publications and highlighted emerging research connections across diagnostic, predictive, and clinical application domains (Fig. 6).

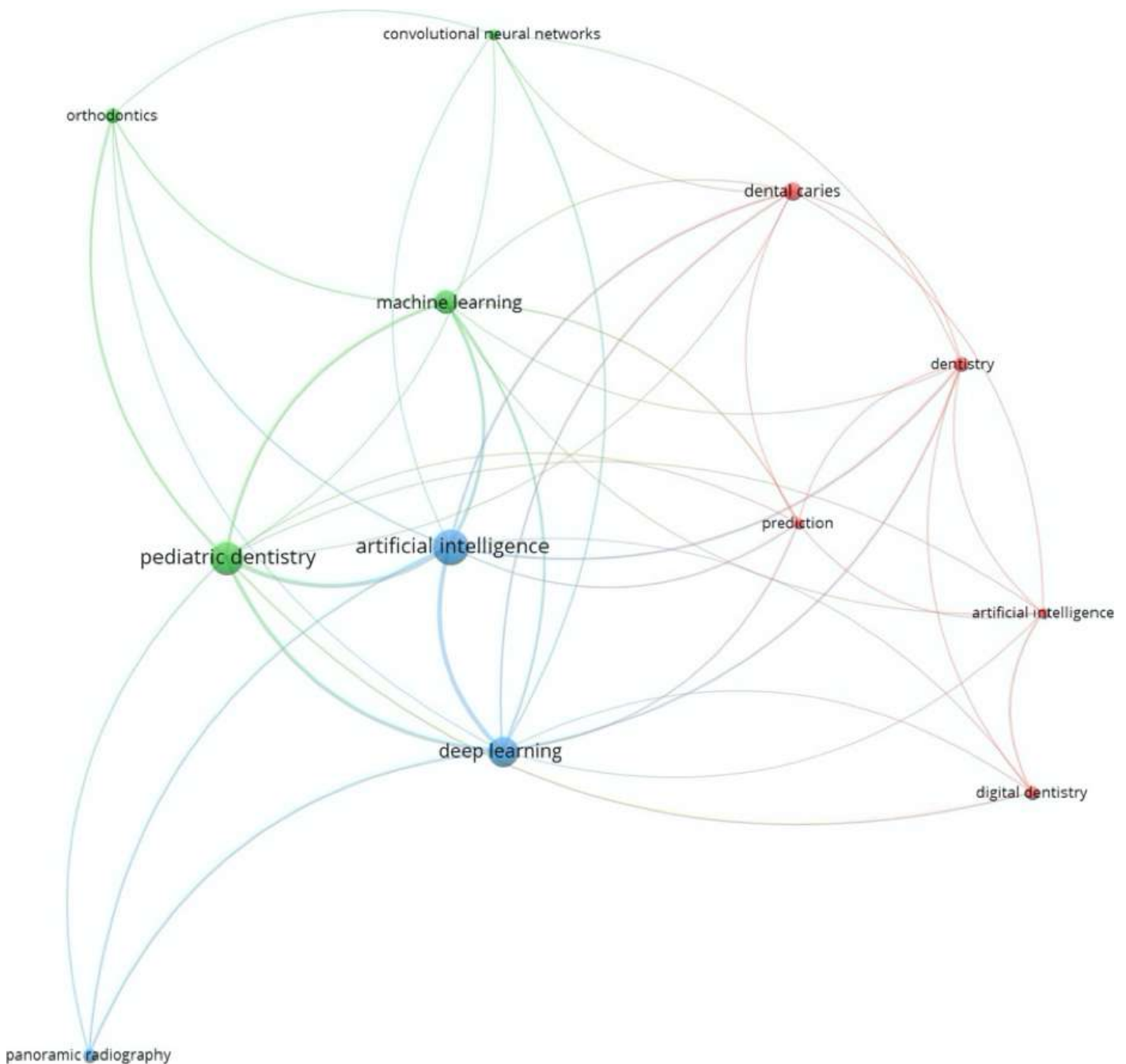
## Discussion

This bibliometric study provides a pediatric dentistry-focused synthesis of research on AI, offering a structured overview of publication trends, thematic orientations, and collaboration patterns within this emerging field. The findings demonstrate a clear increase in AI-related

publications after 2020, particularly during 2023 and 2024, consistent with the broader expansion of AI research across healthcare disciplines.<sup>[12,13]</sup> Despite this quantitative growth, the existing literature remains largely concentrated on diagnostic applications, indicating that AI integration in pediatric dentistry is still in an early stage of development.

At the global level, India, Turkey, and the United States emerged as the most productive contributors, consistent with previous bibliometric analyses in dentistry and pediatric research.<sup>[16–19]</sup> Earlier studies have emphasized the predominance of observational and diagnostically oriented research designs in pediatric dental journals.<sup>[19]</sup> The present findings align with this pattern, highlighting a relatively limited number of experimental or clinically interventional studies, which may constrain the translation of AI technologies into routine pediatric dental practice.

Authorship and collaboration analyses further revealed that scientific output is concentrated among a relatively

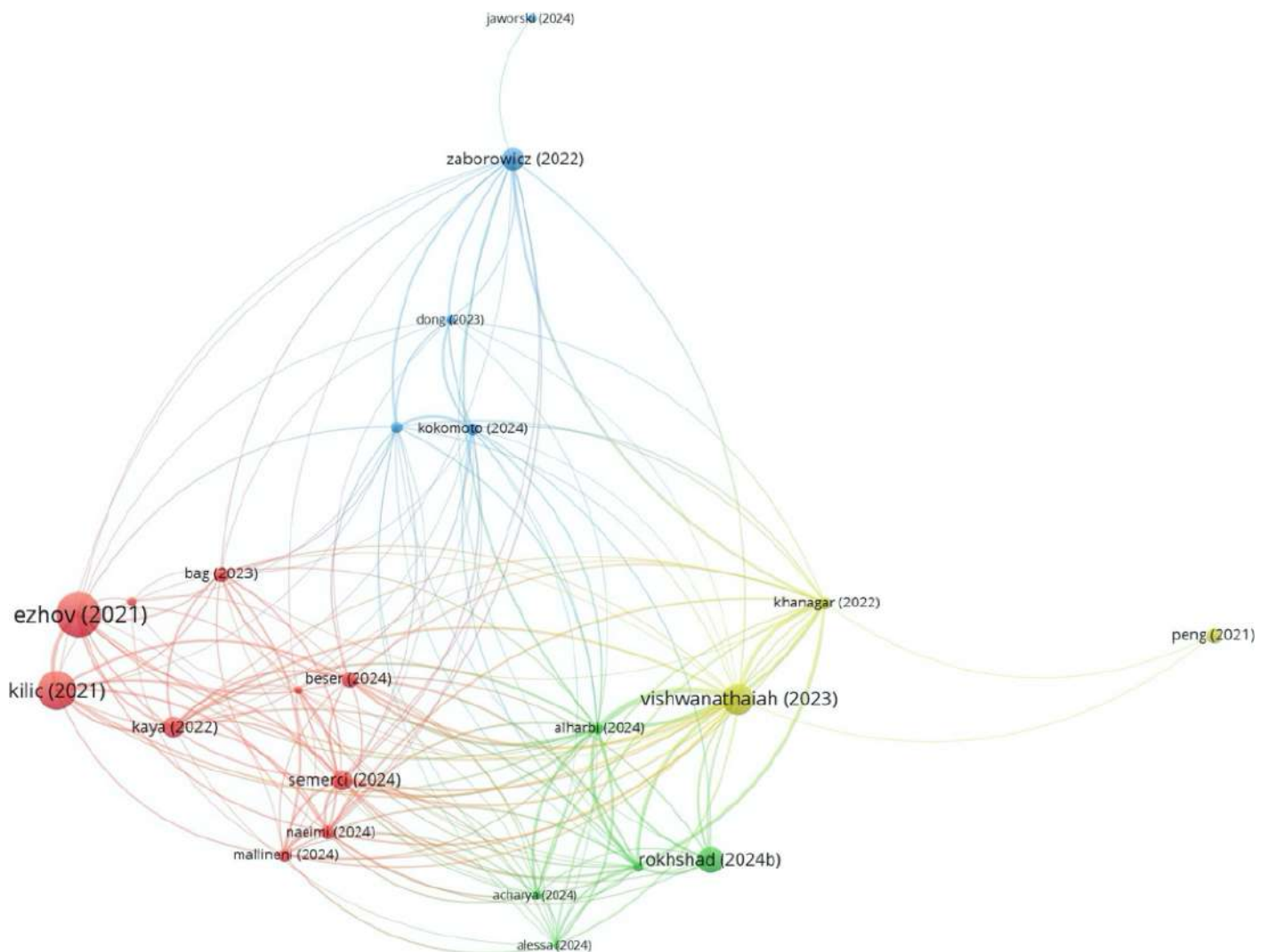


**Figure 5.** Network visualization of keyword co-occurrence in artificial intelligence research in pediatric dentistry. The network was generated using VOSviewer software based on bibliometric data retrieved from the Scopus database.

small number of researchers and research groups. Similar patterns have been reported in bibliometric studies of AI applications in endodontics and implantology, where early growth phases are characterized by fragmented collaboration networks and a predominance of methodological or review-based publications.<sup>[17,18]</sup> These findings suggest that AI research in pediatric dentistry is still consolidating its intellectual and collaborative structure.

Thematic analysis based on keyword co-occurrence

demonstrated a strong emphasis on diagnostic imaging, caries detection, and orthodontic growth assessment. In contrast, topics related to preventive dentistry, behavior management, and patient-reported outcomes were comparatively underrepresented. This imbalance is particularly important given the central role of prevention and behavior guidance in pediatric dental care. Previous bibliometric research has highlighted increasing interest in non-pharmacological behavior management approaches,



**Figure 6.** Network visualization of author relationships in artificial intelligence research in pediatric dentistry. The network was generated using VOSviewer software based on bibliometric data retrieved from the Scopus database.

including virtual reality and digital distraction techniques.<sup>[19]</sup> The limited representation of such themes in AI-focused pediatric dental research indicates important opportunities for future investigations.

From a clinical perspective, AI applications in pediatric dentistry have the potential to extend beyond diagnostic support toward improving behavioral management, early risk prediction, and personalized preventive strategies. AI-assisted tools may contribute to identifying dental anxiety, optimizing communication approaches, and enhancing patient cooperation during clinical procedures.<sup>[20,21]</sup> However, the implementation of such technologies in pediatric settings should remain clinician-guided and ethically grounded, with particular attention to transparency, data security, and the protection of children's rights.

From a national perspective, Türkiye has emerged as a notable contributor to AI research in pediatric dentistry, particularly in diagnostic imaging and radiographic analysis.<sup>[7,8,20–23]</sup> Nevertheless, the relatively limited extent of international collaboration observed in this study is consistent with findings from broader AI bibliometric analyses in healthcare.<sup>[13]</sup> Strengthening cross-regional and interdisciplinary collaborations may therefore be essential to improving methodological rigor, increasing data diversity, and enhancing the clinical applicability of AI-based approaches.

Citation and bibliographic coupling analyses highlighted the continued influence of early foundational studies alongside the emergence of more recent publications forming distinct research fronts.<sup>[24–27]</sup> This pattern reflects the gradual consolidation of AI research in pediatric

dentistry while also indicating a relatively narrow base of highly influential studies compared with more established AI domains in dentistry. In addition, the relatively limited number of publications identified in this study may reflect both the emerging nature of AI applications in pediatric dentistry and the specificity of the search strategy.

The strengths of this study include its systematic bibliometric design and the combined use of two major international databases. Nevertheless, several limitations should be acknowledged. Limitations related to database selection, language bias, and the dynamic nature of citation metrics may have influenced the results. In addition, the search strategy was restricted to selected AI-related terms such as “artificial intelligence,” “machine learning,” and “deep learning.” Although these terms capture a substantial portion of the literature, additional keywords such as “neural networks,” “computer vision,” or “AI-assisted diagnosis” were not included, and some relevant studies may therefore have been missed.

Future bibliometric analyses may benefit from a broader and more inclusive search strategy. Furthermore, future research should prioritize preventive and behavioral applications, incorporate ethical considerations such as transparency and data privacy, and promote collaborative, clinically oriented study designs. Addressing these aspects may facilitate a more balanced, responsible, and patient-centered integration of AI technologies into pediatric dental practice.

## Conclusion

This bibliometric analysis demonstrates that research on AI in pediatric dentistry has grown rapidly in recent years, led primarily by contributions from India, Turkey, and the United States. The literature remains predominantly focused on diagnostic applications, whereas preventive and behavioral domains continue to be comparatively underrepresented. Addressing these gaps through clinically validated, ethically informed, and interdisciplinary research efforts may facilitate a more balanced integration of AI into pediatric dental practice and support the development of evidence-based clinical and educational frameworks.

**Ethics Committee Approval:** As this study relied exclusively on publicly available bibliographic metadata and did not involve human or animal participants, ethical approval was not required.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

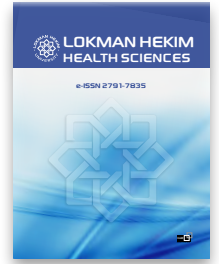
**Use of AI for Writing Assistance:** The author used artificial intelligence (AI)-based tools exclusively for language refinement and readability enhancement. All conceptualization, data analysis, and scientific interpretations were performed solely by the author. Full responsibility for the content and integrity of the manuscript rests with the author.

**Peer-review:** Double blind peer-reviewed.

## References

- Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res* 2020;99(7):769-74. [CrossRef]
- Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *J Dent* 2018;77:106-11. [CrossRef]
- Ekert T, Krois J, Meinhold L, Elhennawy K, Emara R, Golla T, et al. Deep learning for the radiographic detection of apical lesions. *J Endod* 2019;45:917-22.e5. [CrossRef]
- Abuabara A, do Nascimento TVPM, Trentini SM, Costa Gonçalves AM, Hueb de Menezes-Oliveira MA, et al. Evaluating the accuracy of generative artificial intelligence models in dental age estimation based on the Demirjian's method. *Front Dent Med* 2025;6:1634006. [CrossRef]
- Mohajeri A, Schlaud S, Spector S, Hung M. Machine learning for child oral health: a scoping review. *Appl Sci* 2024;14(23):11073. [CrossRef]
- Acharya S, Godhi BS, Saxena V, Assiry AA, Alessa NA, Dawasaz AA, et al. Role of artificial intelligence in behavior management of pediatric dental patients-a mini review. *J Clin Pediatr Dent* 2024;48:24-30. [CrossRef]
- Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in pulp therapy for immature permanent teeth: a comparison of ChatGPT-4 omni, deepseek, and gemini advanced in accuracy, completeness, response time, and readability. *J Endod* 2025;51(11):1675-84. [CrossRef]
- Akgül EN, Gucyetmez Topal B. Detection of Dens Invaginatus on Panoramic Radiographs Using Deep Learning Algorithms. *Int J Paediatr Dent* 2026;36:59-66. [CrossRef]
- Guo X, Shao Y. AI-driven dynamic orthodontic treatment management: personalized progress tracking and adjustments-a narrative review. *Front Dent Med* 2025;6:1612441. [CrossRef]
- Bansal RK, Arya A, Singh B, Singla M, Gupta S. Role of artificial intelligence and machine learning in conservative dentistry and endodontics: A Review. *Cureus* 2025;17(7):e88515. [CrossRef]
- Hasan F, Tantawi ME, Haque F, Foláyan MO, Virtanen JI. Early childhood caries risk prediction using machine learning approaches in Bangladesh. *BMC Oral Health* 2025;25(1):49. [CrossRef]
- Aydın N. Bibliyometric analysis of artificial intelligence research in Türkiye. *BAUNSOBED* 2024;26(3):512-28. [Article in Turkish]

13. Gündüz A, Eren M. The role of artificial intelligence in health sciences: a bibliometric evaluation. *Sağlık Akademisyenleri Dergisi* 2024;11(2):95-105. [Article in Turkish]
14. Royle P, Kandala NB, Barnard K, Waugh N. Bibliometrics of systematic reviews: analysis of citation rates and journal impact factors. *Syst Rev* 2013;2:74. [CrossRef]
15. Park I, Yoon B. Identifying promising research frontiers of pattern recognition through bibliometric analysis. *Sustainability* 2018;10(11):4055. [CrossRef]
16. Cagetti MG, Campus G. The future of pediatric dentistry is now. *Children (Basel)* 2023;10(1):97. [CrossRef]
17. Atalay B. Artificial intelligence in endodontics: a bibliometric analysis. *Turk Klin J Dent Sci* 2025;31(1):45-53. [CrossRef]
18. Erdem R, Genç YS, Yalınız G, Gülşen İT. Dental implantolojide yapay zeka kullanımının bibliyometrik analizi. *Aydin Dental Journal* 2024;10(2):155-67. [Article in Turkish] [CrossRef]
19. Shehani A F, Ponraj S, Ramar K, A VS, S R, J G. Non-pharmacological behavior management techniques in pediatric dentistry: a bibliometric analysis. *Cureus* 2023;15(7):e41329. [CrossRef]
20. Bağ İ, Bilgir E, Bayrakdar İŞ, Baydar O, Atak FM, Çelik Ö, et al. An artificial intelligence study: automatic description of anatomic landmarks on panoramic radiographs in the pediatric population. *BMC Oral Health* 2023;23(1):764. [CrossRef]
21. Kaya E, Gunec HG, Aydin KC, Urkmez ES, Duranay R, Ates HF. A deep learning approach to permanent tooth germ detection on pediatric panoramic radiographs. *Imaging Sci Dent* 2022;52(3):275-81. [CrossRef]
22. Orhan AI, Bezgin T, Orhan K. Digital dentistry applications in pediatric dentistry. In: Delantoni A, Orhan K, editors. *Digital Dentistry*. Cham: Springer; 2024. p. 223-36. [CrossRef]
23. Beser B, Reis T, Berber MN, Topaloglu E, Gungor E, Kılıc MC, et al. YOLO-V5 based deep learning approach for tooth detection and segmentation on pediatric panoramic radiographs in mixed dentition. *BMC Med Imaging* 2024;24:172. [CrossRef]
24. Kwan JC, Fritz PC, Nguyen TT. Artificial intelligence in dentistry: current applications and future perspective. In: Choi AH, editor. *Handbook of tissue reconstruction and regeneration*. Singapore: Springer; 2026. [CrossRef]
25. Moore MM, Slonimsky E, Long AD, Sze RW, Iyer RS. Machine learning concepts, concerns and opportunities for a pediatric radiologist. *Pediatric Radiology* 2019;49:509-16. [CrossRef]
26. Rakhmatulla N, Iroda N, Sherzodbek R. Artificial intelligence applications in pediatric orthodontics: a comprehensive review of current technologies and clinical outcomes. *American Journal of Medicine and Medical Sciences* 2025;15(10):3426-31.
27. Al-Khalifa KS, Ahmed WM, Azhari AA, Qaw M, Alsheikh R, Alqudaihi F, et al. The use of artificial intelligence in caries detection: a review. *Bioengineering* 2024;11(9):936. [CrossRef]



# Determining the Artificial Intelligence Literacy and Attitudes toward Artificial Intelligence of Oncology Nurses: A Descriptive-correlational Study

Zeynep Karakuş<sup>1</sup>, Remziye Semerci Şahin<sup>2</sup>, Burcu Babadağ Savaş<sup>3</sup>, Gülcan Bağçivan<sup>4</sup>

<sup>1</sup>Department of Internal Medicine Nursing, Faculty of Nursing, Akdeniz University, Antalya, Türkiye

<sup>2</sup>Department of Pediatric Nursing, Faculty of Nursing, Koç University, İstanbul, Türkiye

<sup>3</sup>Department of Radiotherapy, Hannover Medical School, Hannover, Germany

<sup>4</sup>College of Nursing and Health Sciences, University of Massachusetts Dartmouth, Dartmouth, Massachusetts, USA

## Abstract

**Introduction:** Artificial intelligence (AI) is increasingly shaping oncology nursing practice. This study examined the relationship between AI literacy and attitudes toward AI among oncology nurses in Türkiye and evaluated AI literacy's predictive role for these attitudes.

**Methods:** This descriptive-correlational study was conducted among oncology nurses between August and December 2025. Data were collected using the personal information form, the Artificial Intelligence Literacy Scale (AILS), and the General Attitudes toward Artificial Intelligence Scale (GA AIS) through Google Forms. Data analysis performed using IBM SPSS Statistics (version 28.0).

**Results:** The study included 102 oncology nurses across Türkiye. Significant positive correlations were found between the total AILS score and all subscales. A strong correlation was observed between usage and ethics, while evaluation showed moderate correlations with the other subscales. In regression analysis, the model predicting negative attitudes was not statistically significant ( $F=0.936$ ,  $p=0.483$ ), explaining 6.5% of the variance ( $R^2=0.065$ ); only Ethics showed a significant negative association. The model predicting positive attitudes was statistically significant ( $F=3.425$ ,  $p=0.003$ ), explaining 20% of the variance ( $R^2=0.20$ ); however, none of the AI literacy subscales were significant predictors, and age showed significant negative association.

**Discussion and Conclusion:** It is recommended that nurses should be supported with in-service training to enhance their AI literacy and foster more positive attitudes toward AI and that relevant regulatory frameworks should be strengthened.

**Keywords:** Artificial Intelligence attitudes; Artificial intelligence literacy; Descriptive study; Nursing; Oncology

Artificial intelligence (AI) is becoming a part of our daily lives faster than other technologies.<sup>[1,2]</sup> AI is currently one of the most powerful transformative forces,

particularly in the healthcare sector.<sup>[3]</sup> In healthcare, AI is used in a wide range of applications, encompassing both basic and advanced care and treatment.<sup>[4]</sup> As global

**Cite this article as:** Karakuş Z, Semerci Şahin R, Babadağ Savaş B, Bağçivan G. Determining the Artificial Intelligence Literacy and Attitudes toward Artificial Intelligence of Oncology Nurses: A Descriptive-correlational Study. Lokman Hekim Health Sci 2026;6(2):237–245.

**Correspondence:** Remziye Semerci Şahin, PhD. Koç Üniversitesi Hemşirelik Fakültesi, İstanbul, Türkiye

**E-mail:** rsemerci@ku.edu.tr **Submitted:** 05.02.2026 **Revised:** 29.04.2026 **Accepted:** 11.05.2026 **Available Online:** 02.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



demographic trends lead to increased prevalence of chronic diseases, patient loads and nurse responsibilities are increasing, and the potential of AI to enhance personalized patient care and workflow efficiency is widely recognized.<sup>[3,5]</sup>

Oncology nurses play many important roles, including evaluating individuals diagnosed with cancer, facilitating diagnostic and therapeutic procedures, and providing personalized, holistic care.<sup>[3,6]</sup> Nurses are collaborating with AI systems, and this is not passive adoption but rather taking an active role in interpreting AI-generated insights, integrating AI into clinical decision-making processes, and providing critical human oversight.<sup>[7]</sup> While AI is transforming nurses' roles in the care of patients with cancer, adopting AI into patient care is closely related to factors such as nurses' knowledge, acceptance, and attitudes.<sup>[4,8,9]</sup> The technology acceptance model (TAM) suggests that perceived ease of use also influences perceived usefulness and that these two perceptions shape individuals' attitudes toward technology, which in turn affects their intention to use technology and their actual usage behavior.<sup>[10]</sup> In this context, the utility and ease of use of AI technologies can be influenced by AI literacy, and nurses must possess a specific level of expertise and a positive attitude to effectively employ them.<sup>[11]</sup>

The objective of AI literacy is to equip individuals with new competencies and methods to engage in a digital society.<sup>[12,13]</sup> To better understand the impact of AI literacy on nursing practices, it is important to examine this concept alongside attitudes. Understanding oncology nurses' attitudes toward AI applications is crucial for successfully integrating AI into clinical settings.<sup>[8]</sup> The growing use of AI in nursing requires oncology nurses to be AI-literate and to maintain a positive attitude.<sup>[9]</sup> In the literature, there are studies on Turkish nurses' AI literacy and attitudes toward AI.<sup>[7,11,14-17]</sup> While some of these studies included nurses working in various wards,<sup>[7,16,17]</sup> others were conducted on specific groups working in a particular ward, such as pediatric or perioperative nurses, etc.<sup>[11,14,15]</sup> Despite the expansion of literature on AI in nursing, there is a gap in research on AI literacy and attitudes among oncology nurses. The study aimed to determine the relationship between AI literacy and attitudes toward AI of oncology nurses in Türkiye and to examine the predictive power of AI literacy on AI attitudes. It is anticipated that the findings of this study will contribute to determining AI literacy and oncology nurses' attitudes toward AI and to developing strategies to strengthen their competencies in AI use.

## Materials and Methods

### Study Design

The study used a descriptive-correlational design. The Strengthening the Reporting of Observational Studies in Epidemiology checklist was used to report the study<sup>[18]</sup> (Appendix 1).

Research questions were

1. What is the level of AI literacy among oncology nurses in Türkiye?
2. What is the general attitude level toward AI among oncology nurses in Türkiye?
3. Is there a significant relationship between the AI literacy levels of oncology nurses and their general attitudes toward AI?
4. Do nurses' AI literacy significantly predict their general attitudes toward AI?

### Setting and Sample

This study was conducted between August and December 2025 with oncology nurses throughout Türkiye. The sample was recruited through convenience sampling. It consisted of oncology nurses who met the inclusion criteria and were actively working in oncology settings such as medical oncology or hematology wards and chemotherapy units. The inclusion criteria for participants were as follows: (a) Actively working in the field of oncology in Türkiye (e.g., medical oncology, chemotherapy unit, hematology-oncology wards, etc.); (b) ability to use a smartphone, social media, or internet to complete the online survey; and (c) voluntary agreement to participate in the study. The exclusion criteria were as follows: (a) Individuals who did not have a digital device with internet access, such as a smartphone, tablet or computer; (b) who did not have digital literacy to fill out data collection tools completely using online survey platforms; (c) the individuals wishes to terminate their voluntary participation at any stage of the research process; (d) individuals who filled out the data collection tools incompletely (20% and above).

Since no similar study in the literature addresses sample size calculation, it was assumed that a moderate correlation ( $r=0.30$ ) would exist between the scores of the Artificial Intelligence Literacy Scale (AILS) and the General Attitudes toward Artificial Intelligence Scale (GAAIS) used in the study. In line with this assumption, the minimum sample size was calculated as 138 nurses using G\*Power 3.1.9.4 software with 95% power and 5% type I error ( $\alpha$ ).<sup>[19]</sup> Considering a possible 10% data loss, the target sample

size was determined as 152. A post hoc power analysis was conducted based on the observed correlation between the positive attitude subscale and the total scale score ( $r=0.309$ ,  $n=102$ ,  $\alpha=0.05$ , two-tailed). This post hoc power analysis ( $r=0.309$ ,  $n=102$ ,  $\alpha=0.05$ , two-tailed) indicated 99% statistical power. However, because post hoc power analyses based on observed correlations have inherent limitations and do not directly reflect the power of regression models, the statistical power of regression analyses was also evaluated separately. Based on the observed model effect sizes, the achieved power was approximately 0.70 for the model predicting a negative attitude and 0.80 for the model predicting a positive attitude.

### Data Collection Tools

Data were collected using a personal information form, the AILS, and the GAAIS.

#### Personal Information Form

The personal information form, prepared by researchers to assess individual characteristics, consists of 10 items such as age, gender, education level, marital status, and the use of AI applications.

#### The AILS

The AILS consists of 12 items and four sub-scales (Awareness, Usage, Evaluation, and Ethics).<sup>[1,20]</sup> Items are scored using a seven-point Likert scale (1=strongly disagree and 7=strongly agree). The lowest score on the AILS is 12, and the highest score is 84, with higher scores indicating greater AI literacy. In the Turkish validity and reliability study of the AILS, the overall Cronbach's  $\alpha$  was 0.85.<sup>[1]</sup> In our study, the total Cronbach's  $\alpha$  value was 0.89, 0.69 for awareness, 0.54 for usage, 0.82 for evaluation, and 0.70 for ethics.

#### The GAAIS

The GAAIS consists of two sub-scales: Negative attitudes toward AI and positive attitudes toward AI. Items are scored using a five-point Likert scale.<sup>[21,22]</sup> In the Turkish validity and reliability study of the scale, Cronbach's  $\alpha$  was 0.82 for positive attitudes toward AI and 0.84 for negative attitudes toward AI.<sup>[22]</sup> In our study, Cronbach's  $\alpha$  was 0.89 for positive attitudes toward AI and 0.82 for negative attitudes toward AI.

### Data Collection

The Google Forms link containing the data collection tools was shared through email or message with hospital administrators, chief nurses, and unit nurses to ensure it reached oncology nurses throughout Türkiye. In addition,

the managers of professional associations, such as the Turkish Oncology Nursing Society, were contacted to ensure that the survey link was announced in oncology nurses' WhatsApp groups. Participants who voluntarily agreed to participate were directed to the data collection tools, and the survey was terminated for those who did not indicate consent. The data collection tools comprised 42 items, with an average completion time of 15 min.

### Statistical Analysis

Statistical analyses were performed using IBM Statistical Package for the Social Sciences Statistics (Version 28.0, IBM Corp, Armonk, NY). Descriptive statistics were calculated and presented as means and standard deviations for continuous variables and as frequencies and percentages for categorical variables. The normality of the data distribution was assessed using the Kolmogorov–Smirnov test, as well as skewness and kurtosis values; values between  $-2$  and  $+2$  were considered acceptable. A Pearson correlation analysis was conducted to examine relationships between the AILS and GAAIS subscales. Multiple linear regression analyses were performed to evaluate the predictive effects of AILS subscales on both negative attitude and positive attitude toward AI. Before regression analyses, multicollinearity among independent variables was assessed using tolerance values and the variance inflation factor (VIF). Variables were retained in the regression models if they met the following criteria: VIF values  $<10$ , tolerance values  $>0.20$ , and condition index values below 15, indicating the absence of multicollinearity. All statistical tests were evaluated at the 95% confidence interval, and  $p<0.05$  was considered statistically significant. Demographic variables were not included in the regression models because the primary aim of the analysis was to examine the predictive relationship between AI literacy dimensions and attitudes toward AI. In addition, given the relatively modest sample size, limiting the number of predictors in the model helped reduce the risk of model overfitting.

### Ethical Approval and Considerations

Ethical approval was obtained from the Koç University Social and Human Sciences Ethics Committee (Date: 24.07.2025, Decision no: 2025.312.IRB3.115) to conduct the research. Before participating in the study, each participant was informed about the study, provided consent, and completed the data collection tools through Google Forms. The participant's confidentiality was ensured with anonymization. The study was conducted in accordance with the Declaration of Helsinki.

## Results

The study was completed with 102 oncology nurses who met the eligibility criteria. The mean age of the nurses was  $37.02 \pm 8.65$ , the nursing experience mean was  $15.41 \pm 9.24$ , and the oncology nurses' experience was  $8.07 \pm 6.79$ . Most of the participants were female (96.08%). Regarding perceived information technology use, nearly half of the participants rated it as medium (47.06%), followed by good (44.12%). Regarding the use of AI in clinical practice, most participants reported not using AI (73.53%), whereas approximately one-quarter reported using it (26.47%) (Table 1). The most frequently reported purpose was the development of patient education materials (56.3%) (Fig. 1).

Based on the GAAIS score, positive attitude showed a small-to-moderate positive correlation with usage ( $r=0.28$ ,  $p=0.042$ ) and a moderate correlation with the AILS total scale score ( $r=0.33$ ,  $p=0.010$ ). Awareness demonstrated moderate positive correlations with usage ( $r=0.38$ ,  $p=0.001$ ), evaluation ( $r=0.53$ ,  $p=0.011$ ), and a strong positive correlation with the total scale ( $r=0.79$ ,  $p<0.001$ ). Usage was moderately associated with evaluation ( $r=0.37$ ,  $p=0.002$ ) and strongly associated with ethics ( $r=0.52$ ,  $p<0.001$ ) and the total scale ( $r=0.75$ ,  $p<0.001$ ). In addition, evaluation was moderately correlated with ethics ( $r=0.48$ ,  $p<0.001$ ) and strongly correlated with the total scale ( $r=0.77$ ,  $p<0.001$ ). A strong positive correlation was also observed between ethics and the total scale ( $r=0.73$ ,  $p<0.001$ ). No other correlations reached statistical significance after correction for multiple comparisons (Table 2).

The linear regression model examining the predictive effects of awareness, usage, evaluation, ethics, and demographic variables (age, nursing experience, and oncology experience) on negative attitude was not statistically significant ( $F=0.936$ ,  $p=0.483$ ), explaining a small proportion of the variance ( $R^2=0.065$ ). This indicates that the predictors did not significantly explain variation in negative attitude. Only Ethics demonstrated a statistically significant association with negative attitude ( $B=-0.54$ , standard error [SE]=0.27,  $p=0.047$ ), suggesting that higher ethics scores may be associated with lower negative attitudes. However, this finding should be interpreted with caution, given the non-significant overall model. Overall, the findings suggest limited explanatory power of the model, and the results should be considered exploratory, with no causal inferences drawn (Table 3).

The linear regression model examining the predictive effects of awareness, usage, evaluation, ethics, and the

**Table 1.** Sociodemographic characteristics of oncology nurses (n=102)

Variables	n	%
Age (years), Mean±SD	37.02±8.65	
Nursing experience (years), Mean±SD	15.41±9.24	
Oncology experience (years), Mean±SD	8.07±6.79	
Gender		
Female	98	96.08
Male	4	3.92
Education level		
High school	12	11.76
Associate degree	14	13.73
Bachelor	52	50.98
Master	21	20.59
Doctorate	3	2.94
Marital status		
Married	69	67.65
Single	33	32.35
Income-expense status		
Income is less than expenses	35	34.31
Income is equal to expenses	59	57.85
Income is greater than expenses	8	7.84
Perceived level of information technology use		
Poor	2	1.96
Medium	48	47.06
Good	45	44.12
Advance	7	6.86
Use of AI in clinical practice		
No	75	73.53
Yes	27	26.47

SD: Standard deviation; AI: Artificial intelligence.

same demographic variables on positive attitude was statistically significant ( $F=3.425$ ,  $p=0.003$ ), explaining 20% of the variance ( $R^2=0.20$ ). This indicates that the model significantly explains variation in positive attitudes. However, none of the primary predictors were statistically significant ( $p>0.05$ ). Among the demographic variables, only age approached statistical significance ( $B=-0.47$ ,  $SE=0.25$ ,  $p=0.047$ ), suggesting a potential negative association with positive attitude, although this did not reach conventional significance levels. Overall, while the model demonstrated significant explanatory power, no individual variable uniquely accounted for variation in positive attitude, and findings should be interpreted cautiously (Table 4).

**Table 2.** Pearson correlation matrix among GAAIS and AILS

Variables	1	2	3	4	5	6	7
GAAIS							
1. Negative attitude	–						
2. Positive attitude	0.15	–					
AILS							
3. Awareness	0.20	0.23*	–				
4. Usage	-0.03	0.28*	0.38*	–			
5. Evaluation	0.02	0.20	0.53*	0.37*	–		
6. Ethics	-0.16	0.24	0.48*	0.52*	0.4*	–	
7. Total scale	0.03	0.33*	0.79*	0.75*	0.77*	0.73*	–

\*Indicates p<0.05. AILS: Artificial intelligence literacy scale; GAAIS: General attitudes toward artificial intelligence scale.

**Table 3.** Results for linear regression with awareness, usage, evaluation, and ethics predicting a negative attitude

Variables	B	SE	95.00% CI	β	t	p
(Intercept)	25.11	3.51	(18.15, 32.0)	0.00	7.16	<0.001
Awareness	0.12	0.18	(0.07, 0.81)	0.26	2.39	0.504
Usage	-0.05	0.23	(-0.50, 0.40)	-0.03	-0.23	0.818
Evaluation	0.10	0.18	(-0.25, 0.45)	0.06	0.57	0.572
Ethics	-0.54	0.27	(-1.09, -0.002)	-0.24	-1.99	<b>0.047</b>
Age	-0.08	0.19	(-0.45, 0.29)	-0.15	-0.43	0.662
Nursing experience (years)	0.05	0.17	(-0.29, 0.39)	0.10	0.30	0.763
Oncology experience (years)	0.13	0.10	(-0.06, 0.33)	0.19	1.32	0.190

SE: Standard error; CI: confidence interval; F=0.936; p=0.483; R<sup>2</sup>=0.065.

**Table 4.** Results for linear regression with awareness, usage, evaluation, and ethics predicting positive attitude

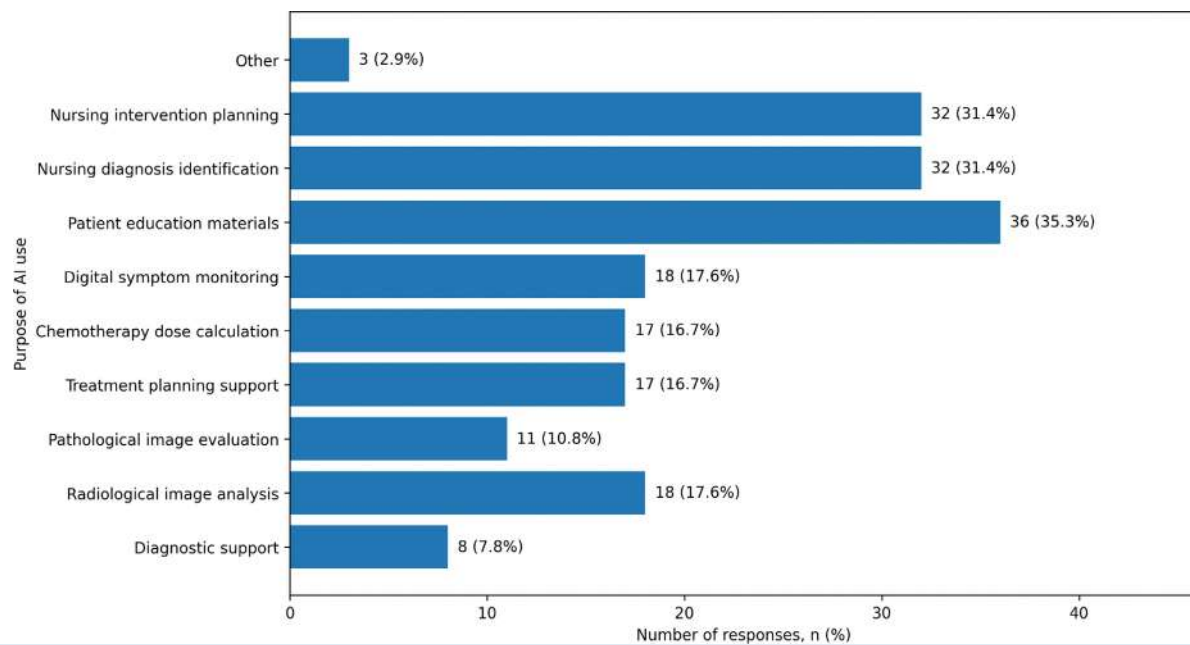
Variables	B	SE	95.00% CI	β	t	p
(Intercept)	25.86	4.96	(16.02, 35.70)	0.00	5.22	<0.001
Awareness	0.35	0.26	(-0.17, 0.87)	0.14	1.34	0.183
Usage	0.45	0.32	(-0.19, 1.08)	0.17	1.41	0.163
Evaluation	0.12	0.25	(-0.37, 0.61)	0.05	0.49	0.628
Ethics	0.28	0.39	(-0.48, 1.05)	0.09	0.73	0.467
Age	-0.47	0.25	(-0.93, 0.01)	-0.59	-1.93	0.047
Nursing experience (years)	0.31	0.23	(-0.15, 0.75)	0.41	1.35	0.181
Oncology experience (years)	-0.12	0.13	(-0.38, 0.15)	-0.12	-0.88	0.380

SE: Standard error; CI: confidence interval; F=3,425; p=0.003; R<sup>2</sup>=0.20.

## Discussion

This study investigated the relationship between AI literacy and oncology nurses' attitudes toward AI in Türkiye and examined AI literacy's predictive power for AI attitudes. The results indicated that most nurses did not use AI in clinical practice and rated their perceived level of information technology use as moderate. It is also found that nurses use AI to develop patient education materials, identify

nursing diagnoses, and plan nursing interventions. These results align with the literature.<sup>[9,17,23]</sup> Similar to our results, a recent systematic review showed that nurses have a poor level of AI knowledge.<sup>[9]</sup> Previous studies also showed that nurses use ChatGPT and AI programs to create nursing care plans, promote their personal development, and calculate medication dosages.<sup>[17,23]</sup> A study concluded that approximately half of the nurses use AI for patient



**Figure 1.** Purposes of AI use.

monitoring, route planning, and nursing documentation.<sup>[24]</sup> Similar to our results, a study examining nurses' attitudes toward AI in Türkiye indicated that, on average, half of the nurses were unaware of AI programs, and the majority of nurses did not use them.<sup>[17]</sup> In a cross-sectional study, nurses' attitudes toward AI technologies were also found to be moderate.<sup>[23]</sup> In this digital era, where AI is rapidly becoming widespread in healthcare delivery, knowledge and perceptions regarding AI are critical determinants of adoption, as stated in TAM.<sup>[10,25]</sup> It is recommended that nurses' use of AI should be increased and that practical training should be organized to positively influence their attitudes toward AI technologies in patient care.

In this study, the awareness, usage, evaluation, and ethics sub-scales of AILS showed strong positive correlations with the total AILS score. A strong correlation was found between ethics and usage and a moderate correlation between evaluation and the other three sub-dimensions. This may indicate that nurses who use and analytically evaluate AI have become more sensitive to ethical issues and have increased their awareness of them. However, the regression models demonstrated low explanatory power, accounting for only 9% of the variance in negative attitudes and 11% in positive attitudes. These findings suggest that although awareness, use, evaluation, and ethics are related to AI literacy, they account for only a limited proportion of nurses' attitudes toward AI. Based on these results, it is recommended that healthcare institutions, policymakers, and administrators plan

technology-based training programs that holistically address awareness, evaluation, use, and ethics issues to increase nurses' AI literacy. At the same time, future studies should examine broader individual and organizational determinants to better explain variations in both positive and negative attitudes toward AI.

The findings regarding negative attitudes toward AI should be interpreted cautiously, as the overall regression model was not statistically significant and explained only a small proportion of the variance in negative attitudes. Accordingly, the results of this model should be considered exploratory rather than confirmatory. Although the ethics sub-scale score showed a significant association within the model, this finding should be interpreted with caution and not considered a robust or independent predictor of negative attitudes toward AI. Rather, it may indicate a limited association, given the model's low explanatory power. While no causal inferences can be drawn, one possible explanation is that nurses with greater ethical awareness may be more sensitive to the potential risks of AI use, including issues of accountability, data privacy, and patient safety. This heightened sensitivity may be related to more cautious or critical attitudes toward AI technologies. Previous studies have reported that nurses' negative perceptions of AI are associated with concerns about malfunctions and errors.<sup>[24]</sup> Moreover, nurses may perceive themselves as responsible for protecting patient privacy and express heightened ethical sensitivity in technology-integrated

healthcare environments.<sup>[26]</sup> AI-nursing collaboration has the potential to improve patient care, but ethical responsibilities and accountability issues remain critical considerations.<sup>[27]</sup> Given the non-significant model, these interpretations should be viewed as tentative. Therefore, future research is needed to identify additional individual, professional, and organizational factors that may better explain negative attitudes toward AI in nursing practice.

The regression analysis regarding positive attitudes toward AI revealed that the model, including the variables awareness, usage, evaluation, and ethics, was statistically significant but explained only a limited proportion of the variance in positive attitudes. However, because none of the individual predictors reached statistical significance, the regression coefficients were not interpreted individually. Our findings suggest that this pattern may indicate shared variance among the predictors, indicating that the variables capture related aspects of a broader construct and therefore contribute jointly to the model, while their unique contributions remain limited. Such situations may arise when conceptually related variables overlap in what they measure, reducing the statistical power to detect independent effects of each predictor. In addition, the relatively modest sample size may have limited the ability to detect small individual effects within the regression model. Measurement-related factors, such as conceptual proximity between the subscales and potential overlap in item content, may also have contributed to this pattern. An effective adoption and use of new technologies, which support from colleagues and organizations, are the most important factors in fostering positive experiences with AI.<sup>[28]</sup> A study emphasized that nursing students and nurses should be equipped with the knowledge and skills to safely integrate AI-assisted health technologies into practice to support care.<sup>[29]</sup> In another study, a strong positive correlation was found between knowledge and positive attitudes toward AI in healthcare, and age, gender, educational background, years of experience, attending conferences, and individuals learning about AI through research articles/journal websites were determinants for a positive attitude toward AI.<sup>[30]</sup> These findings suggest that positive attitudes toward AI may emerge from the combined influence of multiple interrelated competencies rather than a single dominant factor. Consistent with TAM, attitudes toward AI are linked to perceived ease of use and perceived usefulness. Specifically, the observed association between use (as an indicator of adoption behavior) and evaluation and positive attitudes shows

that nurses develop more positive attitudes when they perceive AI as both useful and more applicable in clinical practice. Therefore, multiple factors need to be addressed to foster a positive attitude toward AI use among nurses.

This study has several limitations. First, although the priori power analysis indicated a target sample size of 152 participants, the study was completed with 102 oncology nurses who met the eligibility criteria. Therefore, the final sample remained below the initially planned size, which should be acknowledged as a methodological limitation. In addition, although a post hoc power analysis was performed, this approach has inherent limitations because it is based on observed effect sizes. Moreover, the smaller-than-planned sample size may have reduced the regression analyses' statistical power, potentially limiting the ability to detect significant relationships. As statistical power decreases with smaller sample sizes, the risk of Type II error increases, meaning that some true associations may not be identified. In addition, smaller samples may lead to less stable regression estimates, which should be considered when interpreting the non-significant findings of the regression models. In addition, the use of convenience sampling and voluntary participation may have introduced selection bias, as individuals who were more interested in or familiar with AI may have been more likely to participate. This may have influenced the study findings and should be considered when interpreting the results. Participating oncology nurses may have differed in their opportunities to access and use AI technologies, which may have influenced their responses. Finally, the internal consistency of the usage subscale of the AILS yielded a Cronbach's  $\alpha$  of 0.54 in the present sample. Although the total scale demonstrated excellent reliability ( $\alpha=0.89$ ), the low alpha for this subscale suggests that its items may not have been perceived as measuring a unified construct within this sample. Findings derived from this subscale should therefore be interpreted with caution. Future studies are encouraged to examine the psychometric performance of this subscale across different nursing populations.

## Conclusion

The study found that most oncology nurses did not use AI in clinical practice. Ethical risks were prominent in negative attitudes toward AI, while positive attitudes, although significant in a model including awareness, usage, evaluation, and ethics variables, explained only a limited portion of the variance. Given the cross-sectional design and the models' limited explanatory power, these findings should be interpreted with caution. Rather than providing

definitive recommendations, the results suggest that nurses may benefit from in-service and practical training in AI use and that supportive strategies at the institutional level could be considered to enhance AI literacy and foster more positive attitudes toward AI. Similarly, addressing ethical concerns related to AI use in patient care may be important, and policymakers and healthcare administrators could consider developing appropriate guidelines and frameworks.

**Ethics Committee Approval:** This study was approved by the Koç University Social and Human Sciences Ethics Committee (Date: 24.07.2025, Decision no: 2025.312.IRB3.115).

**Informed Consent:** Written informed consent was obtained.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The authors declared that this study has received no financial support.

**Use of AI for Writing Assistance:** The generative AI tools like QuillBot and Grammarly were used in a limited capacity during the writing process to check grammar and improve sentence clarity.

**Authorship Contributions:** Concept: ZK, RSŞ, BBS, GB; Design: ZK, RSŞ, BBS, GB; Supervision: RSŞ, GB; Data collection and/or processing: ZK, RSŞ, BBS, GB; Analysis and/or interpretation: RSŞ; Literature review: ZK, BBS; Writing: ZK, RSŞ, BBS; Critical review: ZK, RSŞ, BBS, GB.

**Peer-review:** Double blind peer-reviewed.

## References

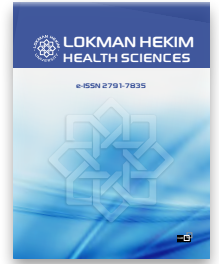
- Çelebi C, Yılmaz F, Demir U, Karakuş F. Artificial intelligence literacy: An adaptation study. *ITALL* 2023;4(2):291-306.
- Temple S, Rowbottom C, Simpson J. Patient views on the implementation of artificial intelligence in radiotherapy. *Radiography* 2023;29:S112-S6. [\[CrossRef\]](#)
- Hassanein S, El Arab, RA, Abdibo A, Abu-Mahfouz MS, Gaballah MKF, Seweid MM, et al. Artificial intelligence in nursing: An integrative review of clinical and operational impacts. *Front Digit Health* 2025;7:1552372. [\[CrossRef\]](#)
- Hamedani Z, Moradi M, Kalroozi F, Manafi Anari A, Jalalifar E, Ansari A, et al. Evaluation of acceptance, attitude, and knowledge towards artificial intelligence and its application from the point of view of physicians and nurses: A provincial survey study in Iran: A cross-sectional descriptive-analytical study. *Health Sci Reports* 2023;6(9):e1543. [\[CrossRef\]](#)
- Alqaissi N, Qtait M. Knowledge, attitudes, practices, and barriers regarding the integration of artificial intelligence in nursing and health sciences education: a systematic review. *SAGE Open Nurs* 2025;2:11:23779608251374185. [\[CrossRef\]](#)
- O'Connor S, Vercell A, Wong D, Yorke J, Fallatah FA, Cave L, et al. The application and use of artificial intelligence in cancer nursing: A systematic review. *Eur J Oncol Nurs* 2024;68:102510. [\[CrossRef\]](#)
- Erkayıran O, Aslan R. Evaluation of nurses' perceptions and readiness for artificial intelligence integration in healthcare: A cross-sectional study in Turkey. *J Adv Nurs* 2025;0:1-13. [\[CrossRef\]](#)
- AL-Olaimat K, Salameh B, Alqadi RA, Alruwaili A, Hakami M, ALanazi HH, et al. Readiness and acceptance of nursing students regarding AI-based health care technology on the training of nursing skills in Saudi Arabia: Cross-sectional study. *JMIR Nurs* 2025;(1):e71653. [\[CrossRef\]](#)
- Sandanasamy S, McFarlane P, Okamoto Y, Couper AL. Nurses' knowledge and attitudes towards artificial intelligence and related factors: A systematic review. *J Nurs Rep Clin Pract* 2025;3(5):486-93. [\[CrossRef\]](#)
- Davis FD. Perceived Usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 1989;13(3):319-40. [\[CrossRef\]](#)
- Subaşı DÖ, Sümengen AA, Semerci R, Şimşek E, Çakır GN, Temizsoy. Paediatric nurses' perspectives on artificial intelligence applications: A cross-sectional study of concerns, literacy levels and attitudes. *J Adv Nurs* 2025;81(3):1353-63. [\[CrossRef\]](#)
- Yılmaz FGK, Yılmaz R. Adaptation of Artificial Intelligence Literacy Scale into Turkish. *Journal of Information and Communication Technologies* 2023;5(2):172-90. [\[CrossRef\]](#)
- Kimiafar K, Sarbaz M, Tabatabaei SM, Ghaddaripouri K, Mousavi AS, Mehneh MR, et al. Artificial intelligence literacy among healthcare professionals and students: a systematic review. *Front Health Inform* 2023;12:168. [\[CrossRef\]](#)
- Güneş H, Okutan Ş, Dolanbay N, Bostancı ME. Investigation of surgical nurses' anxiety and general attitudes towards artificial intelligence. *J Basic Clin Health Sci* 2025;9(3):583-92. [\[CrossRef\]](#)
- Kahraman H, Akutay S, Kaçmaz HY, Taşci, S. Artificial Intelligence literacy levels of perioperative nurses: the case of Türkiye. *Nurs Health Sci* 2025;27(1)e70059. [\[CrossRef\]](#)
- Nirgiz C, Sarı MK, Çaylı N. The relationship between nurses' anxiety and attitudes towards artificial intelligence and examination of influencing factors. *BMC Nurs* 2026;25(1):122. [\[CrossRef\]](#)
- Tuncer GZ, Tuncer M. Investigation of nurses' general attitudes toward artificial intelligence and their perceptions of ChatGPT usage and influencing factors. *Digit Health* 2024;25:10:20552076241277025. [\[CrossRef\]](#)
- Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Lancet* 2007;370(9596):1453-7. [\[CrossRef\]](#)
- Faul F, Erdfelder E, Lang AG, Buchner A. G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 2007;39(2):175-91. [\[CrossRef\]](#)
- Wang B, Rau PLP, Yuan T. Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behav Inf Technol* 2023;42(9):1324-37. [\[CrossRef\]](#)

21. Schepman A, Rodway P. Initial validation of the general attitudes towards Artificial Intelligence Scale. *Comput Hum Behav Rep* 2020;1:100014. [\[CrossRef\]](#)
22. Kaya F, Aydin F, Schepman A, Rodway P, Yetişensoy O, Kaya MD. The roles of personality traits, AI anxiety, and demographic factors in attitudes toward artificial intelligence. *Int J Hum Comput Interact* 2024;40(2):497-514. [\[CrossRef\]](#)
23. Alruwaili MM, Abuadas FH, Alsadi M, Alruwaili AN, Ramadan OME, Shaban M, et al. Exploring nurses' awareness and attitudes toward artificial intelligence: Implications for nursing practice. *Digit Health* 2024;10:20552076241271803. [\[CrossRef\]](#)
24. Sommer D, Schmidbauer, L, Wahl F. Nurses' perceptions, experience and knowledge regarding artificial intelligence: results from a cross-sectional online survey in Germany. *BMC Nurs* 2024;23(1):205. [\[CrossRef\]](#)
25. El Arab RA, Alshakihs AH, Alabdulwahab SH, Almubarak YS, Alkhalifah SS, Abdrbo A et al. Artificial intelligence in nursing: A systematic review of attitudes, literacy, readiness, and adoption intentions among nursing students and practicing nurses. *Front. Digit. Health* 2025;7:1666005 [\[CrossRef\]](#)
26. Rony MKK, Numan SM, Akter K, Tushar H, Debnath M, Johra FT, et al. Nurses' perspectives on privacy and ethical concerns regarding artificial intelligence adoption in healthcare. *Heliyon* 2024;10(17):36702. [\[CrossRef\]](#)
27. Ceviz A, Özden G. Ethical implications of artificial intelligence in intensive care: A literature review of clinical decision-making and nursing practices. *J Intensive Care Nurs* 2025;29(2):152-60.
28. Konttila J, Siira H, Kyngäs H, Lahtinen M, Elo S, Kääriäinen M, et al. Healthcare professionals' competence in digitalization: A systematic review. *J Clin Nurs* 2019;28(5-6):745-61. [\[CrossRef\]](#)
29. Buchanan C, Howitt ML, Wilson R, Booth RG, Risling T, Bamford M. Predicted influences of artificial intelligence on nursing education: Scoping review. *JMIR Nurs* 2021;4(1):e23933. [\[CrossRef\]](#)
30. Rony MKK, Akter K, Nesa L, Islam MT, Johra FT, Akter F, et al. Healthcare workers' knowledge and attitudes regarding artificial intelligence adoption in healthcare: A cross-sectional study. *Heliyon* 2024;10(23):e40775. [\[CrossRef\]](#)

**Appendix 1. STROBE statement – Checklist of items that should be included in reports of cross-sectional studies**

	Item no	Recommendation	Pages
Title and abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract	1
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found	1
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	1–2
Objectives	3	State specific objectives, including any prespecified hypotheses	2
Methods			
Study design	4	Present key elements of study design early in the paper	2
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	2
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants	2, 3
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	3–4
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	3–4
Bias	9	Describe any efforts to address potential sources of bias	NA
Study size	10	Explain how the study size was arrived at	2–3
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	3
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding	4
		(b) Describe any methods used to examine subgroups and interactions	4
		(c) Explain how missing data were addressed	4
		(d) If applicable, describe analytical methods taking account of sampling strategy	NA
		(e) Describe any sensitivity analyses	NA
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	4
		(b) Give reasons for non-participation at each stage	NA
		(c) Consider use of a flow diagram	NA
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	4–5
		(b) Indicate number of participants with missing data for each variable of interest	NA
Outcome data	15*	Report numbers of outcome events or summary measures	
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	5
		(b) Report category boundaries when continuous variables were categorized	5
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NA
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	NA
Discussion			
Key results	18	Summarise key results with reference to study objectives	6–7
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	7
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	7, 8
Generalisability	21	Discuss the generalisability (external validity) of the study results	7, 8
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	COI disclosure

\*Give information separately for exposed and unexposed groups. An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at <http://www.plosmedicine.org/>, Annals of Internal Medicine at <http://www.annals.org/>, and Epidemiology at <http://www.epidem.com/>). Information on the STROBE Initiative is available at [www.strobe-statement.org](http://www.strobe-statement.org).



## ORIGINAL ARTICLE

# Assessment of Artificial Intelligence Chatbots' Information Quality on Home Dental Bleaching: A Comparative Study

Özlem Özişçi, Elif Irmak Gökmen

Department of Prosthodontics, Süleyman Demirel University Faculty of Dentistry, Isparta, Türkiye

## Abstract

**Introduction:** The purpose of this research was to compare how artificial intelligence chatbots (ChatGPT-3.5, ChatGPT-4, Gemini, and DeepSeek) responded to common patient inquiries regarding home dental bleaching, with an emphasis on quality, accuracy, clarity, and practical applicability.

**Methods:** Forty patient-oriented questions identified using the AlsoAsked tool, which extracts Google "People Also Ask" data, were categorized into seven thematic domains and submitted individually to each chatbot in separate sessions. Responses were independently scored by two evaluators using the global quality scale (GQS), accuracy of information index (AOI), and patient education materials assessment tool for printed materials. Response times were recorded in seconds. Statistical analyses included the Kruskal–Wallis test, Bonferroni-adjusted pairwise comparisons, and Spearman correlation ( $p < 0.05$ ).

**Results:** ChatGPT-4 and DeepSeek achieved the highest GQS and AOI scores. DeepSeek had the highest actionability score but the longest response time. ChatGPT-3.5 demonstrated moderate performance, while Gemini had the lowest intelligibility and actionability scores.

**Discussion and Conclusion:** Advanced artificial intelligence chatbots can provide high-quality and accurate information on at-home dental bleaching. However, unsupervised use may pose patient safety risks; thus, their deployment should be limited to validated, monitored, and task-specific applications.

**Keywords:** Artificial intelligence; Chatbot; Patient education; Tooth bleaching

Artificial intelligence (AI) includes computational techniques that process large datasets, recognize patterns, and improve task performance over time.<sup>[1]</sup> These approaches from machine learning to deep-learning models enable automated feature extraction, predictive modeling, and decision support in clinical contexts.<sup>[2]</sup> In dentistry, AI applications have enhanced diagnostic consistency in radiographic interpretation, caries detection,

and treatment planning.<sup>[2–5]</sup> Although performance varies with task complexity and data quality, validated models can reliably perform well-defined dental tasks under clinician supervision.<sup>[1,2]</sup>

Chatbots based on large language models (LLMs) extend AI's use by providing scalable, on-demand patient education and guidance.<sup>[6–8]</sup> Beyond patient communication, LLMs support evidence summarization and administrative tasks,

**Cite this article as:** Özişçi Ö, Gökmen El. Assessment of Artificial Intelligence Chatbots' Information Quality on Home Dental Bleaching: A Comparative Study. Lokman Hekim Health Sci 2026;6(2):246–254.

**Correspondence:** Özlem Özişçi, M.D. Süleyman Demirel Üniversitesi, Diş Hekimliği Fakültesi, Protetik Diş Hekimliği Anabilim Dalı, Isparta, Türkiye  
**E-mail:** oslemozisci@gmail.com **Submitted:** 20.10.2025 **Revised:** 28.03.2026 **Accepted:** 16.05.2026 **Available Online:** 02.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



but may also produce inaccurate or fabricated information, posing safety risks if unsupervised.<sup>[9–13]</sup> Therefore, safe use requires task-specific validation and human oversight.

Dental bleaching is among the most common esthetic procedures.<sup>[14,15]</sup> Patients frequently request information about efficacy, safety, and risks.<sup>[14,16]</sup> While dentists conventionally provide such knowledge, AI chatbots are increasingly consulted as alternative sources.<sup>[13,17]</sup> However, their outputs may lack scientific rigor and completeness.<sup>[13,17,18]</sup> Given that bleaching outcomes depend on product choice, patient suitability, and post-treatment care, the accuracy of information is critical.<sup>[16,19]</sup>

Accordingly, a systematic evaluation of AI chatbots' bleaching-related responses is needed. Assessing quality, accuracy, readability, and guideline consistency will clarify current capabilities and limitations.<sup>[13,17]</sup> In this study, we aimed to comparatively evaluate the quality, accuracy, readability, and guideline consistency of responses generated by four AI chatbots to common patient questions on at-home dental bleaching. The null hypothesis of this study was that there would be no statistically significant differences among the evaluated AI chatbots in terms of information quality (Global Quality Scale [GQS]), accuracy of information index (AOI), intelligibility, actionability, and response duration when answering patient-oriented questions about at-home dental bleaching.

## Materials and Methods

### Study Place and Design

This study was conducted at the department of prosthodontics, faculty of dentistry. The research process included question identification, chatbot response collection, and evaluator scoring within the same period.

### Type of Research

This was a cross-sectional, comparative, and descriptive study designed to evaluate and compare the quality of information provided by four AI chatbots on the topic of home dental bleaching.

### Population and Sample

The study population consisted of AI Chatbot-generated responses rather than human participants. Responses from four AI chatbots (ChatGPT-3.5, ChatGPT-4, Gemini, DeepSeek) were compared. A total of 40 patient-oriented questions on at-home bleaching were selected to represent the most frequently searched public inquiries. The questions were generated using the AlsoAsked

platform (<https://alsoasked.com>), which aggregates Google "People Also Ask" queries derived from Google Search results (<https://www.google.com>). Specifically, queries were retrieved by entering the keywords "home dental bleaching" and related search terms into the Google search engine, and the associated "People Also Ask" suggestions were systematically extracted through AlsoAsked on July 10, 2025. Because Google search results are dynamically updated and may vary by user location, time, and search history, the retrieval date and keyword strategy were documented to enhance reproducibility. All extracted questions were recorded verbatim at the time of data collection.

The four AI chatbots evaluated in this study (ChatGPT-3.5, ChatGPT-4, Gemini, and DeepSeek) were selected based on predefined criteria to ensure representativeness and methodological rigor. First, these platforms were among the most widely accessible and publicly used LLM-based chatbots at the time of data collection, reflecting real-world tools that patients are likely to consult for health-related information. Second, the selected models represent different developers and architectural frameworks (OpenAI, Google, and DeepSeek AI), allowing cross-platform comparison across systems trained on distinct datasets and alignment strategies. Third, prior healthcare and dentistry-related comparative studies have evaluated these models, establishing their relevance in clinical and patient education contexts. Including both earlier generation (ChatGPT-3.5) and more advanced models (ChatGPT-4, Gemini, and DeepSeek) also allowed evaluation of potential performance differences across model generations. Therefore, the selection of these four chatbots was intended to ensure a representative, cross-platform, and clinically relevant comparison of widely used LLM-based systems with differing architectures, generations, and prior healthcare applicability.

### Questionnaire Structure

A total of 40 questions on at-home bleaching were categorized into seven domains: General definition and mechanism, eligibility, safety, effectiveness, causes, application, and post-treatment maintenance. Most of the questions were related to safety and effectiveness, highlighting patients' primary concerns regarding the use of whitening products at home, as shown in Table 1.<sup>[20,21]</sup>

### Data Collection

All chatbot responses were collected between July 11 and July 15, 2025, using the publicly accessible

**Table 1.** Classification of home bleaching questions

Category	Questions
General definition and mechanism	<p>What is home bleaching, and how does it work?</p> <p>How does light activation (e.g., LED lights) enhance home bleaching?</p> <p>What is the difference between hydrogen peroxide (HP) and carbamide peroxide (CP) in whitening gels?</p> <p>What does the percentage on whitening gels mean (e.g., 10%, 16%, 22%)?</p> <p>How to make teeth white naturally from yellow?</p>
Eligibility and contraindications	<p>Is home bleaching suitable for everyone?</p> <p>Who should avoid it?</p> <p>Can children or teenagers use home whitening products?</p> <p>Can pregnant or breastfeeding women undergo home bleaching?</p> <p>Can people with braces (orthodontic treatment) use whitening products?</p> <p>Can you bleach rotten teeth?</p>
Safety and side effects	<p>Is it safe to whiten your teeth at home?</p> <p>Is CP safe for teeth?</p> <p>Is it safe to put HP directly on your teeth?</p> <p>Is it good to brush your teeth with HP every day?</p> <p>What percentage of peroxide do dentists use to whiten teeth?</p> <p>Is it safe to use high-concentration gels at home?</p> <p>Is there a risk of over-bleaching?</p> <p>Can teeth whitening crack your teeth?</p> <p>Does enamel grow back?</p> <p>What are the side effects of household bleach?</p> <p>How do I manage gum irritation caused by whitening gel?</p>
Effectiveness and expectations	<p>Is in-office teeth whitening better than take-home?</p> <p>Do crest white strips work?</p> <p>Do whitening strips contain the same active ingredients as whitening gels?</p> <p>How many shades whiter can I expect my teeth to become with home bleaching? Does home bleaching produce the same results for every person?</p> <p>Are natural remedies (e.g., baking soda, coconut oil) as effective as home bleaching products?</p>
Causes and pre-treatment conditions	<p>Why are my teeth so yellow?</p> <p>What chemical whitens teeth?</p>
Application frequency and duration	<p>How often should I whiten my teeth?</p> <p>How many days can you use home bleaching?</p> <p>How long does it take to whiten teeth with 6 HP?</p> <p>How long does it take to see noticeable results?</p> <p>How long does home bleaching last?</p>

web-based interfaces of each platform. All platforms were accessed through their publicly available web interfaces using standard user-level accounts without API access, developer-level permissions, or institutional custom configurations. No premium API integrations, external plugins, or experimental features were enabled during data collection. The versions available on the respective

web platforms at the time of access were used without modification. The following model versions were used during the study: ChatGPT-3.5 (OpenAI, GPT-3.5-turbo), ChatGPT-4 (OpenAI, GPT-4), Google Gemini (Google AI, standard publicly available web version), and DeepSeek (DeepSeek LLM, publicly accessible web interface version available at the time of data collection).

All platforms were accessed under default system configurations. No modifications were made to temperature settings, output length parameters, response tone, or formatting options. Each question was entered in a separate new session to minimize contextual carryover effects. No additional prompts, clarifications, or follow-up instructions were provided. Only the first complete response generated by each chatbot was recorded verbatim and included in the analysis without editing, summarization, or structural modification. This approach was adopted to enhance reproducibility and to reflect typical real-world patient interactions with these systems.

### **Evaluation of Responses**

Two independent evaluators assessed chatbot outputs using:

#### **GQS**

The GQS is a five-point Likert-type instrument used to evaluate the overall quality, flow, comprehensiveness, and usefulness of health-related information. Scores range from 1 (poor quality, incomplete, and not useful) to 5 (excellent quality, well-structured, and highly useful). Higher scores indicate better overall informational quality.

#### **AOI**

A 10-point tool evaluating factual correctness, consistency, and relevance. The AOI assessment was conducted using a predefined evidence-based reference framework. Authoritative professional sources were systematically used as benchmarks, including the American Dental Association guidelines on tooth whitening, the FDI World Dental Federation policy statement on bleaching, and peer-reviewed scientific literature addressing home dental bleaching indications, contraindications, mechanisms, and safety considerations. Before scoring, the evaluators reviewed these reference documents to establish a consensus understanding of accepted clinical standards and evidence-based recommendations. Each chatbot response was then compared against these benchmark sources to determine factual correctness, consistency with established guidelines, clinical appropriateness, and completeness of information. When discrepancies existed between sources, priority was given to the most recent evidence-based clinical guideline. This structured benchmarking approach minimized subjective interpretation and enhanced methodological rigor.

### **Patient Education Materials Assessment Tool (PEMAT-P)**

The PEMAT-P was applied in accordance with the original scoring manual developed by Shoemaker et al.<sup>[22]</sup> Each item within the understandability and actionability domains was independently evaluated by two trained reviewers. Items were scored dichotomously as “agree” (1 point) or “disagree” (0 points), while items deemed “not applicable” were excluded from the denominator, as recommended in the official PEMAT-P instructions. Percentage scores for understandability and actionability were calculated by dividing the total number of items scored as “agree” by the total number of applicable items and multiplying the result by 100. This standardized percentage approach enabled direct comparison across chatbot responses despite minor variations in structural formatting. Before formal scoring, both evaluators conducted a calibration session using pilot responses to ensure consistent interpretation of PEMAT-P criteria. Discrepancies in scoring were resolved through structured consensus discussions. Inter-rater reliability for PEMAT-P demonstrated good-to-excellent agreement.

Inter-rater reliability for PEMAT-P percentage scores was specifically assessed using the intraclass correlation coefficient (ICC) under a two-way random-effects model with absolute agreement. The ICC values for PEMAT-P ranged between 0.81 and 0.89, indicating good-to-excellent agreement between evaluators. Table 2 presents the evaluation criteria applied to score chatbot responses. The GQS was assessed on a five-point scale, with higher values reflecting better quality, flow, and completeness of the information provided. The AOI included five items (factual accuracy, corroboration, consistency, clarity, and relevance), each scored from 0 to 2, yielding a maximum total of 10 points. These measures enabled a structured and objective assessment of the reliability and accuracy of chatbot-generated content. Inter- and intra-rater reliability were determined using Cronbach's  $\alpha$  and ICC, which demonstrated good-to-excellent agreement ( $\alpha=0.83-0.87$ ;  $ICC=0.78-0.94$ ), thereby confirming the robustness of the scoring system, as detailed in Table 2.

To ensure methodological transparency and comparability across models, chatbot outputs were not subjected to any artificial constraints or post-processing. No limits were imposed on response length, formatting style, tone, or depth of explanation beyond the default system configurations of each platform. Temperature parameters, response length settings, and stylistic controls were not manually modified. Each question was entered as

**Table 2.** Modified GQS and AOI

Section	Score/item	Description/definition	Maximum score
GQS (Quality/completeness scale)	1	Poor quality, poor flow of the information, most information missing, not at all useful for clinicians	–
GQS (Quality/completeness scale)	2	Generally poor quality and flow, some information listed but many important topics missing, very limited usefulness for clinicians	–
GQS (Quality/completeness scale)	3	Moderate quality, suboptimal flow, some important information adequately discussed, but others poorly discussed, somewhat useful for clinicians	–
GQS (Quality/completeness scale)	4	Good quality and generally good flow. Most of the relevant information is listed, but some topics are not listed, useful for clinicians	–
GQS (Quality/completeness scale)	5	Excellent quality and flow, very useful for clinicians	–
AOI	Factual accuracy	The response aligns with known facts, data, or established knowledge on the subject	2
AOI	Corroboration	The response is based on evidence from textbooks, studies, or guidelines	2
AOI	Consistency	The response is internally consistent and does not contain contradictory statements	2
AOI	Clarity and specificity	The response is clear and specific, avoiding vague or ambiguous language	2
AOI	Relevance	The response directly addresses and adheres to the question or topic posed	2
AOI	Total AOI score	The sum of all scores	10

GQS: Global Quality Scale; AOI: Accuracy of Information Index.

a standalone prompt in a new session, and no follow-up prompts, clarifications, or iterative refinements were provided. Only the first complete response generated by each model was recorded verbatim. Responses were not edited, summarized, reformatted, or truncated before evaluation. This approach was intentionally adopted to simulate real-world patient interactions, where users typically submit a single question and receive an unmodified response. Although evaluation metrics such as GQS and PEMAT-P may be influenced by response verbosity and structural organization, maintaining default output conditions across all models ensured that comparisons reflected inherent model behavior rather than externally imposed standardization procedures.

### Statistical Analysis

Normality was assessed using the Shapiro–Wilk test. As data were not normally distributed, nonparametric tests were applied. Differences among the four AI models were analyzed using the Kruskal–Wallis test, followed by Bonferroni-adjusted pairwise comparisons where appropriate. Associations between quality metrics were evaluated using Spearman correlation analysis. Multiple

linear regression analysis was performed to identify independent predictors of overall information quality (GQS). A two-sided  $p < 0.05$  was considered statistically significant. All analyses were conducted using IBM Statistical Package for the Social Sciences Statistics for Windows, Version 26.0 (IBM Corp., Armonk, NY, USA).

### Ethical Approval

Ethics committee approval was not required, as the study analyzed secondary, de-identified chatbot responses and did not involve human participants, personal data, or interventions. This decision aligns with institutional policy and international ethical standards (e.g., Declaration of Helsinki).

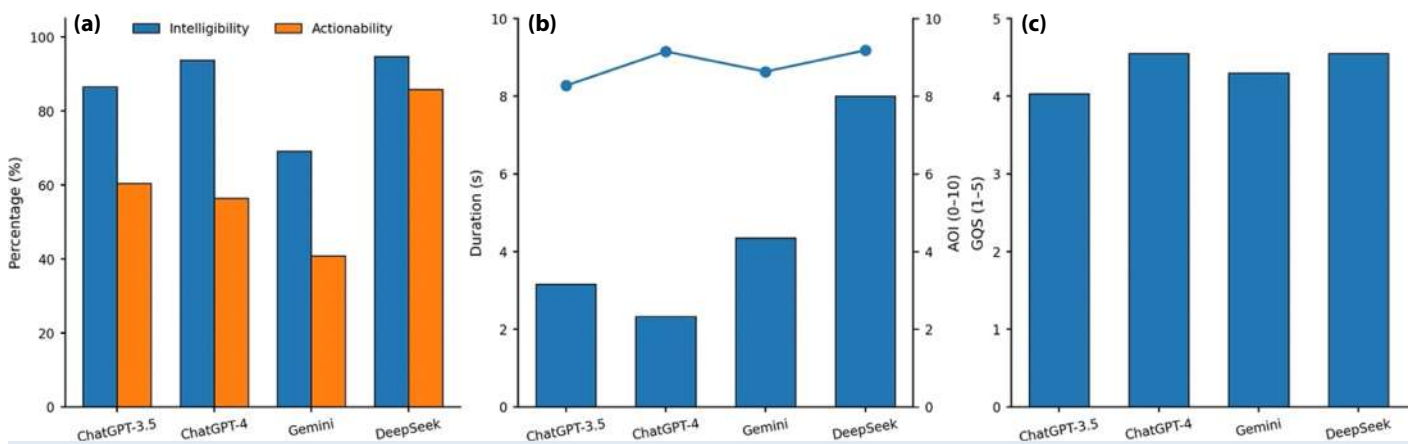
### Results

An evaluation of 40 prompts by two independent raters revealed statistically significant performance differences among the four AI models across all assessed parameters ( $p < 0.001$ ). The comparative results are presented in Table 3 and illustrated in Figure 1. ChatGPT-4 and DeepSeek demonstrated the highest intelligibility scores, whereas Gemini showed the lowest values. Regarding actionability,

**Table 3.** Descriptive statistics and comparison of mean scores for intelligibility, actionability, duration, GQS, and total AOI across artificial intelligence chatbots

Variables	ChatGPT-3.5 (Mean±SD)	ChatGPT-4 (Mean±SD)	Gemini (Mean±SD)	DeepSeek (Mean±SD)	p
Intelligibility score (%)	86.56±16.84 <sup>b</sup>	93.75±6.33 <sup>a</sup>	69.06±12.66 <sup>c</sup>	94.69±6.26 <sup>a</sup>	<0.001*
Actionability (%)	60.42±22.86 <sup>b</sup>	56.41±24.67 <sup>b</sup>	40.83±33.54 <sup>c</sup>	85.84±12.26 <sup>a</sup>	<0.001*
Duration (seconds)	3.16±0.99 <sup>c</sup>	2.32±0.82 <sup>c</sup>	4.35±1.13 <sup>b</sup>	8.00±3.83 <sup>a</sup>	<0.001*
GQS (1–5)	4.03±0.53 <sup>b</sup>	4.55±0.55 <sup>a</sup>	4.30±0.56 <sup>a</sup>	4.55±0.64 <sup>a</sup>	<0.001*
Total AOI score (0–10)	8.27±0.68 <sup>b</sup>	9.15±0.62 <sup>a</sup>	8.63±1.08 <sup>b</sup>	9.18±0.98 <sup>a</sup>	<0.001*

a, b, c: Groups sharing the same letter are not significantly different; different letters indicate significant differences between groups (Bonferroni-adjusted pairwise comparisons,  $p < 0.05$ ); SD: Standard deviation; GQS: Global Quality Scale; AOI: Accuracy of Information Index.



**Figure 1.** Comparison of artificial intelligence chatbot performance across evaluation metrics. (a) Intelligibility and actionability scores (%) presented on a 0–100 scale. (b) Duration (seconds) and accuracy of information index displayed together using a dual-axis format to improve interpretability. (c) Global quality scale scores presented on a 0–5 scale.

DeepSeek achieved the highest mean score, significantly outperforming the other models. ChatGPT-3.5 and ChatGPT-4 exhibited moderate actionability levels, while Gemini had the lowest performance. In terms of response duration, DeepSeek generated significantly longer responses, whereas ChatGPT-4 and ChatGPT-3.5 were the fastest models. For overall quality (GQS), ChatGPT-4 and DeepSeek achieved the highest scores, followed by Gemini, while ChatGPT-3.5 scored comparatively lower. Similarly, for factual accuracy (AOI), ChatGPT-4 and DeepSeek demonstrated superior performance. Overall, ChatGPT-4 and DeepSeek consistently outperformed the other models in intelligibility, overall quality, and factual accuracy. DeepSeek provided the most actionable responses but required longer response times, whereas ChatGPT-4 demonstrated a more balanced profile between quality and efficiency.

Spearman correlation analysis demonstrated a significant positive association between AOI ( $r=0.39$ ;  $p < 0.001$ ) and GQS ( $r=0.75$ ;  $p < 0.001$ ) across models. In contrast, response duration was not significantly correlated with overall quality in most comparisons. This finding indicates that higher

factual accuracy was consistently associated with higher perceived informational quality. In other words, responses that were more aligned with evidence-based standards tended to receive higher overall quality scores. Conversely, the absence of a significant association between response duration and GQS suggests that longer responses did not necessarily translate into higher informational quality. Multiple linear regression analysis was conducted to identify independent predictors of GQS (Table 4). The regression model was statistically significant ( $p < 0.001$ ) and explained approximately 47% of the variance in perceived quality ( $R^2=0.469$ ). Among the evaluated variables, the AOI emerged as the only significant independent predictor of GQS ( $\beta=0.702$ ,  $p < 0.001$ ), whereas understandability, actionability, and response duration did not independently predict overall quality. To further explore the robustness of the regression model, a stepwise regression approach was considered. Given that AOI was the only statistically significant predictor in the full model, it is expected that a stepwise regression model would retain AOI as the sole predictor of GQS. In such a case, a similar coefficient of determination ( $R^2$ ) would likely be obtained, indicating that

**Table 4.** Multiple linear regression analysis identifying independent predictors of overall information quality (GQS)

Predictor	B	SE	$\beta$	p
Understandability	0.003	0.003	0.087	0.215
Actionability	-0.003	0.002	-0.121	0.112
Response duration (seconds)	0.005	0.012	0.025	0.687
AOI	0.456	0.041	0.702	<0.001

Model summary:  $R^2=0.469$ ; Adjusted  $R^2=0.455$ ; F-test  $p<0.001$ . GQS: Global Quality Scale; AOI: Accuracy of Information Index

AOI alone explains a substantial proportion of the variance in overall information quality. These findings further support the central role of factual accuracy in determining perceived information quality. Although intelligibility and actionability contribute to the presentation of information, they did not independently predict overall quality when accuracy was controlled for in the regression model. This suggests that evidence-based correctness remains the primary determinant of high-quality patient-oriented information. Our results clearly show that chatbot performance differs across models, likely reflecting differences in architecture and training approaches.

## Discussion

The null hypothesis of this study stated that there would be no statistically significant differences among the evaluated AI chatbots in terms of information quality (GQS), factual accuracy (AOI), intelligibility, actionability, and response duration when answering patient-oriented questions about at-home dental bleaching. The present findings demonstrated statistically significant inter-model differences across all assessed parameters ( $p<0.001$ ). Accordingly, the null hypothesis was rejected. These results indicate that AI chatbot performance is not uniform and varies meaningfully depending on model architecture, training strategy, and alignment mechanisms, underscoring important mechanistic and clinical considerations regarding their use in dental patient education.<sup>[17,18]</sup>

Performance differences observed among chatbots may be largely attributed to architectural and training-related factors inherent to different model generations. More advanced systems benefit from larger and more diverse training corpora, improved instruction tuning, and refined reinforcement learning with human feedback, which collectively enhance contextual reasoning, internal consistency, and adherence to health-related norms. These mechanisms likely explain the superior balance between factual accuracy and coherence observed in higher-performing models. In contrast, earlier generation models may rely more heavily on surface-level pattern

matching, increasing susceptibility to incomplete explanations or clinically ambiguous guidance, particularly in nuanced topics such as dental bleaching safety and contraindications.<sup>[13,18]</sup>

Another important mechanistic consideration relates to alignment strategies and response generation constraints. Models optimized for conversational fluency may generate longer, more structured responses, which can enhance actionability but also increase the risk of verbosity-driven overgeneralization. Conversely, more conservative alignment approaches may prioritize brevity and factual grounding at the expense of practical guidance. The observed variability in actionability and response duration across models highlights this trade-off and underscores that higher informational quality is not solely dependent on response length but rather on how effectively evidence-based content is translated into patient-understandable recommendations.<sup>[16,19]</sup>

From a clinical perspective, these findings have direct relevance to dental bleaching, where inappropriate self-directed use of whitening agents can lead to adverse outcomes such as tooth sensitivity, soft-tissue irritation, enamel alterations, or unrealistic esthetic expectations. Accurate guidance regarding patient eligibility, product concentration, treatment duration, and post-bleaching care is essential to ensure safety and efficacy. Although AI chatbots may serve as accessible adjuncts for general patient education, their unsupervised use poses potential risks, particularly when responses lack adequate emphasis on contraindications or the need for professional consultation. Therefore, high-performing chatbots should be viewed as supportive informational tools rather than substitutes for clinician-led decision-making.<sup>[23–27]</sup>

The strong association observed between factual accuracy and perceived informational quality further reinforces the central role of evidence-based correctness in AI-generated health content. The regression analysis further confirmed this relationship, as AOI emerged as the only independent predictor of GQS, explaining a substantial proportion of the variance in perceived quality. This statistical finding

strengthens the interpretation that accuracy, rather than stylistic features or response length, is the principal driver of quality perception. This finding suggests that improving the reliability of chatbot outputs requires prioritization of validated clinical knowledge rather than optimization for speed or stylistic complexity. Accordingly, future implementations of AI systems in dentistry should adopt a human-in-the-loop framework, where Chatbot-generated information is integrated into dentist-supervised workflows, such as pre-consultation education or post-treatment instruction reinforcement.<sup>[28]</sup>

Several directions for future research emerge from this study. Longitudinal investigations are needed to evaluate how ongoing model updates influence informational quality over time. Multilingual analyses would further clarify the generalizability of chatbot performance across diverse patient populations. Importantly, experimental studies assessing patient comprehension, behavioral adherence, and clinical outcomes following exposure to AI-generated information would provide critical evidence on real-world effectiveness. In addition, comparative evaluations of domain-specific (fine-tuned dental models vs. general) LLMs may help identify optimal strategies for safely deploying AI tools in clinical dentistry.

Despite these contributions, certain limitations should be acknowledged. The study evaluated chatbot responses at a single time point using standardized prompts, which may not fully capture dynamic conversational interactions. Moreover, outcomes related to patient understanding or behavior were not directly assessed. Nonetheless, by systematically evaluating quality, accuracy, and actionability under controlled conditions, this study provides a robust foundation for future investigations into the responsible integration of AI chatbots into dental patient education.

## Conclusion

In this comparative evaluation of four contemporary AI chatbots, statistically significant differences were identified among the models in terms of information quality (GQS), factual accuracy (AOI), intelligibility, actionability, and response duration. ChatGPT-4 and DeepSeek demonstrated the highest performance in overall quality and accuracy. DeepSeek achieved the highest actionability scores but required longer response times, whereas ChatGPT-4 provided a more time-efficient profile with comparable quality. These findings indicate measurable variability in chatbot performance when responding to patient-oriented questions about at-home dental bleaching.

**Ethics Committee Approval:** Ethics committee approval was not required, as the study analyzed secondary.

**Informed Consent:** Was not required, as the study analyzed secondary.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** No AI tools were used in the generation, analysis, or writing of the scientific content of this manuscript.

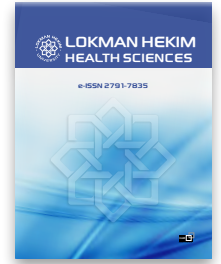
**Authorship Contributions:** Concept: ÖÖ, EIG; Design: ÖÖ, EIG; Supervision: ÖÖ, EIG; Resource: ÖÖ, EIG; Materials: ÖÖ, EIG; Data collection and/or processing: ÖÖ, EIG; Analysis and/or interpretation: ÖÖ, EIG; Literature review: ÖÖ, EIG; Writing: ÖÖ, EIG; Critical review: ÖÖ, EIG.

**Peer-review:** Double blind peer-reviewed.

## References

1. Topol EJ. Deep medicine: How artificial intelligence can make healthcare human again. New York, NY: Basic Books; 2019.
2. Ding H, Wu J, Zhao W, Matinlinna JP, Burrow MF, Tsoi JKH. Artificial intelligence in dentistry-A review. *Front Dent Med* 2023;4:1085251. [\[CrossRef\]](#)
3. Shafi I, Fatima A, Afzal H, Díez IT, Lipari V, Breñosa J, Ashraf I. A Comprehensive review of recent advances in artificial intelligence for dentistry e-health. *Diagnostics (Basel)* 2023;13(13):2196. [\[CrossRef\]](#)
4. Stephan D, Bertsch A, Burwinkel M, Vinayahalingam S, Al-Nawas B, Kämmerer PW, Thiem DG. AI in dental radiology-improving the efficiency of reporting with ChatGPT: comparative study. *J Med Internet Res* 2024;26:e60684. [\[CrossRef\]](#)
5. Farhadi Nia M, Ahmadi M, Irankhah E. Transforming dental diagnostics with artificial intelligence: advanced integration of ChatGPT and large language models for patient care. *Front Dent Med* 2025;5:1456208. [\[CrossRef\]](#)
6. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative ai large language models ChatGPT, google bard, and microsoft bing chat in supporting evidence-based dentistry: comparative mixed methods study. *J Med Internet Res* 2023;25:e51580. [\[CrossRef\]](#)
7. Naik S, Al-Kheraif AA, Vellappally S. Artificial intelligence in dentistry: Assessing the informational quality of YouTube videos. *PLoS One* 2025;20(1):e0316635. [\[CrossRef\]](#)
8. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. *J Esthet Restor Dent* 2023;35(7):1098-102. [\[CrossRef\]](#)
9. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? *J Stomatol Oral Maxillofac Surg* 2023;124(5):101471. [\[CrossRef\]](#)
10. Bender EM, Gebru T, McMillan Major A, Mitchell M. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FACCT '21)* 2021:610-23. [\[CrossRef\]](#)

11. Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58<sup>th</sup> Annual Meeting of the Association for Computational Linguistics. 2020. [\[CrossRef\]](#)
12. Lu X, Zhang R, Wu W, Shang X, Liu M. Relationship between internet health information and patient compliance based on trust: Empirical study. *J Med Internet Res* 2018;20(8):e253. [\[CrossRef\]](#)
13. Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. *Am J Orthod Dentofacial Orthop* 2024;165(6):652-62. [\[CrossRef\]](#)
14. Akarslan ZZ, Sadik B, Erten H, Karabulut E. Dental esthetic satisfaction, received and desired dental treatments for improvement of esthetics. *Indian J Dent Res* 2009;20(2):195-200. [\[CrossRef\]](#)
15. Aldaij M, Alshehri T, Alzeer A, Alfayez A, Aldrees F, Almuhaya S, et al. Patient satisfaction with dental appearance and treatment desire to improve esthetics. *J Oral Health Comm Dent* 2018;12(3):90-5. [\[CrossRef\]](#)
16. Demarco FF, Meireles SS, Masotti AS. Over-the-counter whitening agents: A concise review. *Braz Oral Res*. 2009;23(Suppl 1):64-70. [\[CrossRef\]](#)
17. Guven Y, Ozdemir OT, Kavan MY. Performance of artificial intelligence chatbots in responding to patient queries related to traumatic dental injuries: A comparative study. *Dent Traumatol* 2025;41(3):338-47. [\[CrossRef\]](#)
18. Taymour N, Fouda SM, Abdelrahman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google Gemini large language models in responding to dental implantology inquiries. *J Prosthet Dent* 2025;134(6):2427-34. [\[CrossRef\]](#)
19. Carey CM. Tooth whitening: what we now know. *J Evid Based Dent Pract*. 2014;14 Suppl:70-6. [\[CrossRef\]](#)
20. FDI World Dental Federation. FDI policy statement on dental bleaching materials: adopted by the FDI General Assembly: 17 September 2011 - Mexico City, Mexico. *Int Dent J* 2013;63(1):2-3. [\[CrossRef\]](#)
21. American Dental Association, Council on Scientific Affairs. Tooth whitening/bleaching: Treatment considerations for dentists and their patients. Chicago (IL): American Dental Association; 2010.
22. Shoemaker SJ, Wolf MS, Brach C. Development of the patient education materials assessment tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns*. 2014;96(3):395-403. [\[CrossRef\]](#)
23. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in dentistry: A comprehensive review. *Cureus* 2023;15(4):e38317. [\[CrossRef\]](#)
24. Azadi A, Gorjinejad F, Mohammad-Rahimi H, Tabrizi R, Alam M, Golkar M. Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2024;137(6):587-93. [\[CrossRef\]](#)
25. Babayiğit O, Tasta Eroglu Z, Ozkan Sen D, Ucan Yarkac F. Potential use of ChatGPT for patient information in periodontology: a descriptive pilot study. *Cureus*. 2023;15(11):e48518. [\[CrossRef\]](#)
26. Mahdi SS, Battineni G, Khawaja M, Allana R, Siddiqui MK, Agha D. How does artificial intelligence impact digital healthcare initiatives? A review of AI applications in dental healthcare. *Int J Inf Manag Data Insights* 2023;3(1):100144. [\[CrossRef\]](#)
27. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. *BMC Oral Health* 2025;25(1):573. [\[CrossRef\]](#)
28. Salmanpour F, Akpınar M. Performance of Chat Generative Pretrained Transformer-4.0 in determining labiolingual localization of maxillary impacted canine and presence of resorption in incisors through panoramic radiographs: A retrospective study. *Am J Orthod Dentofacial Orthop* 2025;168(2):220-31. [\[CrossRef\]](#)



# Evaluation of AI Chatbots in Tooth Avulsion Management According to the International Association of Dental Traumatology Guidelines

Merve Özdemir<sup>1</sup>, Esra Yıldırım Manav<sup>2</sup>

<sup>1</sup>Department of Pediatric Dentistry, Faculty of Dentistry, Lokman Hekim University, Ankara, Türkiye

<sup>2</sup>Department of Restorative Dentistry, Faculty of Dentistry, Lokman Hekim University, Ankara, Türkiye

## Abstract

**Introduction:** This study aimed to evaluate the extent to which widely used artificial intelligence (AI)-based chatbots adhere to the 2020 International Association of Dental Traumatology (IADT) guidelines for the management of tooth avulsion and to assess the accuracy of the bibliographic references (i.e., complete citation details including title, authors, journal, year, and DOI) they generate.

**Methods:** This cross-sectional observational study assessed four AI-based chatbots (ChatGPT-5.2, Perplexity AI, Gemini 2.5 Flash, and DeepSeek-v3.2) using ten standardized, clinician-directed avulsion scenarios aligned with the 2020 IADT guidelines. Each scenario was submitted once per chatbot, without iterative prompting, on 3 January 2026. Scenarios varied by extra-oral dry time, storage medium, apex maturity, dentition type, and replantation timing. Responses were evaluated using the 9-item IADT Compliance Index. Bibliographic accuracy was assessed using the reference hallucination score (RHS).

**Results:** No statistically significant difference was observed in overall normalized compliance scores among the chatbots ( $p=0.089$ ). However, significant between-model differences emerged in technically critical domains, including root surface cleaning ( $p=0.017$ ), and splint type and duration ( $p<0.001$ ). ChatGPT-5.2 and Perplexity AI consistently outperformed Gemini 2.5 Flash and DeepSeek-v3.2. Although RHS values did not differ significantly between models ( $p=0.114$ ), all chatbots demonstrated occasional reference hallucinations.

**Discussion and Conclusion:** Performance was higher in simpler scenarios, such as immediate replantation, whereas more complex conditions – particularly prolonged dry time and primary tooth avulsion – showed lower compliance and greater variability. Although chatbots reproduce general principles, limitations restrict reliability; thus, they should be used with clinician supervision.

**Keywords:** Artificial intelligence; Dental trauma; International Association of Dental Traumatology guidelines; Tooth avulsion

Dental avulsion is one of the most severe and time-critical traumatic dental injuries (TDIs), accounting for up to 16% of dental trauma cases in children and adolescents.<sup>[1]</sup>

The prognosis of an avulsed permanent tooth is determined primarily by the duration of extra-oral dry time, the condition of the periodontal ligament (PDL), and the appropriateness

**Cite this article as:** Özdemir M, Yıldırım Manav E. Evaluation of AI Chatbots in Tooth Avulsion Management According to the International Association of Dental Traumatology Guidelines. Lokman Hekim Health Sci 2026;6(2):255–263.

**Correspondence:** Merve Ozdemir, M.D. Lokman Hekim Üniversitesi, Diş Hekimliği Fakültesi, Pediyatrik Diş Hekimliği Bölümü, Ankara, Türkiye

**E-mail:** merveozdemir@lokmanhekim.edu.tr **Submitted:** 16.02.2026 **Revised:** 08.05.2026 **Accepted:** 18.05.2026 **Available Online:** 02.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



of immediate management steps performed at the scene of injury or in the emergency clinic.<sup>[2,3]</sup> Even minor deviations from established protocols, such as using an inappropriate storage medium or delaying replantation, can lead to ankylosis-related replacement resorption and eventual tooth loss. Furthermore, errors in other key management steps – such as inappropriate replantation decisions, incorrect splint duration, or suboptimal antibiotic recommendations – may further compromise periodontal healing, increase the risk of infection or inflammatory root resorption, and negatively affect long-term tooth survival.<sup>[4]</sup>

To ensure standardized, evidence-based management, the International Association of Dental Traumatology (IADT) published comprehensive guidelines in 2020 for both permanent and primary dentitions.<sup>[5,6]</sup> Management is influenced by factors such as apex maturity, extent of PDL damage, and appropriate follow-up for detecting complications. Despite the availability of these guidelines, adherence in clinical practice remains inconsistent. Studies indicate that many general practitioners and emergency physicians are unfamiliar with less common scenarios, such as open-apex avulsion, delayed replantation, or management of contaminated root surfaces.<sup>[7,8]</sup> In addition, real-time emergencies may limit access to clinical protocols, specialist consultation, or institutional resources, particularly in resource-limited settings. Challenges such as determining appropriate splinting duration, assessing tetanus status, and distinguishing between permanent and primary dentition further increase the risk of error.

In recent years, artificial intelligence (AI) chatbots have emerged as tools to address this accessibility gap. With retrieval-augmented generation (RAG) and access to academic databases, they can provide real-time, guideline-based recommendations and bibliographic references. Surveys suggest that more than half of clinicians use AI for decision support, particularly in time-sensitive trauma settings.<sup>[9–11]</sup> Although IADT guidelines are publicly available, accessing them during emergencies is often impractical. In contrast, AI chatbots offer rapid, on-demand responses, making them attractive for acute decision support.<sup>[9,10]</sup> However, their outputs remain unregulated and variable, and it is unclear whether they provide consistent recommendations across diverse avulsion scenarios, including variations in dry time, storage conditions, apex maturity, dentition type, and follow-up requirements.

One emerging concern is reference hallucination, in which AI systems generate citations that appear valid but do not correspond to real publications. Recent evaluations have shown hallucination rates ranging from 30% to 60% in

medical and dental AI-generated reference lists, posing risks for misinformation, incorrect clinical decision-making, and compromised academic integrity.<sup>[12–14]</sup> Ensuring accurate citation of the 2020 IADT guidelines is especially important because management differs substantially depending on factors such as extra-oral time thresholds ( $\leq 30$  min vs.  $\geq 90$  min), apex maturity, and appropriate follow-up intervals.

Recent studies have demonstrated that AI-based chatbots are increasingly being applied across various dental domains, including diagnosis, treatment planning, and patient education.<sup>[15–17]</sup> Previous studies have shown that AI-based systems can generally reproduce broad management principles for TDIs, including avulsion.<sup>[18–20]</sup> However, these studies have primarily focused on overall accuracy rather than compliance with guideline-sensitive, time-dependent clinical steps. Key aspects emphasized in the 2020 IADT guidelines – such as root surface management after prolonged dry time, socket preparation, and splinting protocols – have not been systematically evaluated across diverse, clinically realistic scenarios.

In addition, although some AI tools provide bibliographic references, the accuracy of these citations, particularly regarding the 2020 IADT avulsion guidelines, remains largely unexplored.

This study therefore aimed to evaluate the ability of AI chatbots to follow IADT-recommended avulsion management by applying a standardized prompt encompassing all essential guideline components across diverse clinical scenarios. The null hypothesis was that there would be no significant difference among the evaluated AI chatbots in terms of their compliance with the 2020 IADT avulsion guidelines and the accuracy of their cited references.

## Materials and Methods

### Study Design

This cross-sectional observational study evaluated the adherence of four AI-based chatbots to the 2020 IADT guidelines when responding to clinician-directed avulsion management questions. No human subjects or biological samples were involved; therefore, ethics approval was not required.

The study assessed four widely used English-language AI-based chatbots equipped with real-time literature retrieval systems:

- ChatGPT-5.2 (OpenAI, USA)
- Perplexity AI (Perplexity Inc., USA)
- Gemini 2.5 Flash (Google DeepMind, UK)
- DeepSeek-v3.2 (DeepSeek Lab, China).

## Question Preparation Process

A total of 10 clinically realistic avulsion scenarios were created in alignment with the IADT Guidelines by two dental traumatology specialists with at least 5 years of clinical experience (Supplementary Material 1).

To reflect age-dependent differences in root development, open-apex scenarios were designed to represent 7-year-old children, whereas closed-apex scenarios represented 12-year-old patients. The scenarios incorporated variations in extra-oral dry time (30 vs. 90 min), storage medium (milk for 30 or 90 min), apex maturity (open vs. closed), primary tooth avulsion, and immediate replantation at the accident site.

The final scenario list consisted of:

1. Closed apex – 30 min dry (12 years)
2. Open apex – 30 min dry (7 years)
3. Closed apex – 90 min dry (12 years)
4. Open apex – 90 min dry (7 years)
5. Closed apex – milk 30 min (12 years)
6. Open apex – milk 30 min (7 years)
7. Closed apex – milk 90 min (12 years)
8. Open apex – milk 90 min (7 years)
9. Avulsed primary maxillary incisor
10. Immediately replanted the tooth at the accident site (12 years).

Following each clinical scenario, all chatbots were consistently asked the standardized fixed question: "How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, web link, and DOI." This approach was used to ensure methodological standardization and comparability across all scenarios.

## Question Redirection Process

To minimize potential biases and eliminate memory effects, each clinical scenario was presented in a separate chat session. Thus, ten independent sessions were created for each chatbot, with one scenario asked per session. Each scenario and prompt pair was submitted individually to each chatbot. Before initiating each session, the chat history was cleared, the "delete our chat history" command was executed when applicable, and any remaining messages were manually removed. In addition, the browser cache, cookies, and browsing history were cleared before each session to prevent any carryover effects.

All chatbot interactions were performed using publicly accessible web-based user interfaces (ChatGPT through the OpenAI platform, Perplexity AI through perplexity.ai, Gemini through gemini.google.com, and DeepSeek through chat.deepseek.com), rather than application programming interfaces, under free-tier user access conditions. Model versions were identified based on the labels displayed in the user interface at the time of access. Exact backend version identifiers are not publicly disclosed by the providers; therefore, model identification relied on publicly visible version information. No custom system prompts, developer instructions, or role-based priming were applied. Default system settings were used for all models, and no adjustable generation parameters (e.g., temperature, top-p sampling, or maximum token limits) were modified. Real-time literature retrieval or browsing features (i.e., RAG), when available, were not manually activated, deactivated, or modified, and all models were evaluated under their default system configurations. No plugins, developer modes, or auxiliary tools were activated during data collection. Each prompt was submitted once per scenario, and no follow-up questions or iterative refinements were performed. All scenarios were submitted on the same day (January 03, 2026) to minimize temporal variability. All responses were exported as plain text, anonymized, and coded for evaluation. Each chatbot received the same ten prompts in identical order and formatting, and all prompts were submitted verbatim without modification. No cross-session information sharing occurred between queries. All interactions were carried out by a single experienced investigator (MO) to ensure standardization and consistency.

## Evaluation Process

Although the chatbots were only asked a single clinician-directed question per scenario, their responses were independently evaluated using a predefined 9-item IADT Compliance Index derived from the 2020 IADT guidelines. The evaluation focused on the presence and accuracy of the following management components: (1) Indication for replantation, (2) root surface cleaning, (3) socket preparation, (4) endodontic timing and intracanal medicaments, (5) systemic antibiotic therapy, (6) tetanus prophylaxis, (7) type and duration of splinting, (8) post-operative patient instructions, and (9) clinical and radiographic follow-up schedule. Items not applicable to a given scenario (e.g., replantation for primary teeth) were excluded from the denominator. A compliance percentage was calculated for each response.

## Reference Verification (Reference Hallucination Score [RHS] Framework)

The RHS is a structured framework developed to systematically quantify the severity of hallucinated citations in AI-generated content and was applied in the present study to evaluate each bibliographic reference generated by the chatbots. The term “bibliographic reference” refers to the complete citation details, including title, authors, journal name, year of publication, volume, issue, page numbers, web link, and DOI. Each reference was systematically evaluated across seven predefined bibliographic identifiers, including the accuracy of the title, authors’ names, journal name, publication year, digital object identifier (DOI), and web link (URL), as well as its relevance to the dental trauma topic.<sup>[12]</sup> Each major hallucination (e.g., incorrect or missing title, author list, journal, or DOI) received 2 points, while minor hallucinations (e.g., wrong year, invalid link, or irrelevant topic) received 1 point. Thus, the total RHS per reference ranged from 0 (fully accurate) to 11 (completely hallucinated). In addition, all clinical response components defined within the 9-item IADT Compliance Index, as well as all bibliographic references generated by the chatbots, were independently evaluated by the same two calibrated dental traumatology experts, who were blinded to chatbot identities throughout the assessment process. Clinical items were scored dichotomously as correctly addressed or incorrect/missing, while references were assessed according to predefined RHS criteria. In cases of disagreement, responses and references were re-evaluated, and consensus was achieved through discussion. Interrater reliability was assessed using the intraclass correlation coefficient (ICC), demonstrating excellent agreement for both clinical scoring (ICC=0.93) and RHS assessment (ICC=0.89).

## Statistical Analysis

Statistical analyses were performed using IBM Statistical Package for the Social Sciences Statistics (version 27; IBM Corp., Armonk, NY, USA). Descriptive statistics were calculated for all variables. Total clinical adherence scores were derived from a 9-item IADT-based index and converted to percentage values (0–100). Given that all chatbot models were evaluated using identical clinical scenarios, the data were treated as repeated measures. The Shapiro–Wilk test indicated non-normal distribution of total scores. Accordingly, non-parametric tests appropriate for repeated-measures designs were applied. Comparisons of total clinical adherence scores across chatbot models were performed using the Friedman test. For individual

**Table 1.** Comparison of total clinical adherence scores across chatbot models

Chatbot models	Mean±SD (%)	p
ChatGPT-5.2	70.6±15.7	0.089
Perplexity AI (%)	72.7±14.0	
Gemini 2.5 Flash	75.0±19.5	
DeepSeek-v3.2	60.0±13.0	

Values are presented as mean±standard deviation (SD) percentages based on a 9-item IADT clinical adherence index. Total scores were normalized to a 0–100 scale. Between-model comparisons were performed using the Friedman test. A  $p < 0.05$  was considered statistically significant.

clinical criteria, responses were dichotomized as correct or incorrect and expressed as percentages. Between-model comparisons for these binary variables were conducted using Cochran’s Q test. For criteria demonstrating statistically significant overall differences, post hoc pairwise comparisons were performed using McNemar tests with Bonferroni correction to account for multiple comparisons. RHS values were compared across the four chatbot models using the Kruskal–Wallis test. A two-tailed  $p < 0.05$  was considered statistically significant.

## Results

Comparison of percentage-based total scores across chatbot models showed no statistically significant difference among the models ( $p=0.089$ ), indicating comparable overall clinical performance. In contrast, statistically significant differences emerged in several technical domains requiring precise interpretation of the IADT guidelines. Significant between-group differences were observed for root surface cleaning ( $p=0.017$ ). The greatest divergence among chatbot models was identified in the splint type and duration criterion, which demonstrated a highly significant difference ( $p < 0.001$ ). Mean clinical adherence scores for each chatbot model are presented in Table 1, whereas detailed item-level performance across all evaluated IADT-based criteria is shown in Table 2.

Post hoc pairwise comparisons were performed for clinical criteria demonstrating significant overall differences using McNemar tests with Bonferroni correction (adjusted significance level  $p < 0.008$ ). For root surface cleaning, ChatGPT-5.2 and Perplexity AI showed higher accuracy compared to DeepSeek-v3.2, with differences reaching statistical significance after correction ( $p < 0.008$ ). No statistically significant difference was observed between ChatGPT-5.2 and Perplexity AI. Comparisons involving Gemini 2.5 Flash showed a trend

**Table 2.** Comparison of chatbot performance across IADT-based criteria

Outcome/Criterion	ChatGPT-5.2 (%)	Perplexity AI (%)	Gemini 2.5 Flash (%)	DeepSeek-v3.2 (%)	p	Effect size
Indication for replantation	100.0	100.0	90.0	90.0	0.572	0.11
Root surface cleaning	75.0	75.0	37.5	0.0	0.017	0.42
Replantation site preparation	44.4	100.0	77.8	55.6	0.530	0.28
Endodontic considerations	55.6	66.7	88.9	66.7	0.438	0.15
Antibiotic recommendation	80.0	70.0	90.0	90.0	0.194	0.22
Tetanus prophylaxis	80.0	80.0	90.0	80.0	0.875	0.04
Splint type and duration	88.9	0.0	88.9	55.6	0.001	0.58
Post-operative care	30.0	80.0	40.0	40.0	0.690	0.26
Follow-up schedule	80.0	80.0	70.0	80.0	0.912	0.03

Values represent percentages of correct responses for each clinical criterion. Between-model comparisons were performed using Cochran's Q test. A  $p < 0.05$  was considered statistically significant. Effect size: Kendall's W interpreted as: 0.1–0.3 (Small), 0.3–0.5 (Moderate),  $> 0.5$  (Large). IADT: International Association of Dental Traumatology.

toward lower performance; however, these differences did not consistently reach the adjusted significance threshold. The most pronounced between-model differences were observed for splint type and duration. Both ChatGPT-5.2 and Gemini 2.5 Flash demonstrated significantly higher accuracy than Perplexity AI ( $p < 0.008$ ). No statistically significant difference was detected between ChatGPT-5.2 and Gemini 2.5 Flash. Comparisons involving DeepSeek-v3.2 indicated lower performance but did not consistently reach statistical significance after Bonferroni correction. The comparative performance of the chatbot models across these discriminative criteria is shown in Appendix Figure 1.

Follow-up performance differed across chatbot models. Correct reporting of the recommended clinical and radiographic follow-up schedule was observed in 80% of responses generated by ChatGPT-5.2, Perplexity AI, and DeepSeek-v3.2, whereas Gemini 2.5 Flash demonstrated a lower accuracy of 70%. Although overall compliance scores were comparable, these findings indicate that follow-up recommendations constitute a domain in which clinically relevant discrepancies persist across models, particularly in scenarios requiring structured long-term monitoring. However, the difference in follow-up schedule accuracy among chatbot models did not reach statistical significance ( $p = 0.912$ ).

RHS values across chatbot models are presented in Table 3. Although ChatGPT-5.2 and Perplexity AI exhibited wider score distributions with occasional high outliers, no statistically significant difference in RHS values was observed among ChatGPT-5.2, Perplexity AI, Gemini 2.5 Flash, and DeepSeek-v3.2 ( $H = 5.95$ ,  $p = 0.114$ ).

**Table 3.** Comparison of RHS values across chatbot models

Chatbot	n	Mean±SD	Median	Min–Max
ChatGPT-5.2	10	1.6 ± 2.46	1	0–8
DeepSeek-v3.2	10	1.6 ± 0.84	2	0–2
Gemini 2.5 Flash	10	1.6 ± 0.84	2	0–2
Perplexity AI	10	1.1 ± 2.51	0	0–8

Values are presented as mean±standard deviation (SD), median, and minimum–maximum. RHS values were compared across chatbot models using the Kruskal–Wallis test. A  $p < 0.05$  was considered statistically significant. RHS: Reference hallucination score.

## Discussion

The present study evaluated the extent to which four widely used AI-based chatbots comply with the 2020 IADT avulsion guidelines when responding to standardized, clinician-directed scenarios. In accordance with the study hypothesis, no statistically significant difference was observed in overall normalized performance scores among the chatbot models; therefore, the null hypothesis was accepted. However, meaningful disparities emerged in specific guideline-sensitive domains. These findings suggest that overall performance metrics may mask potential deficiencies in technically demanding aspects of avulsion management that could influence prognosis.<sup>[2–6]</sup>

The absence of statistically significant differences in total performance scores suggests that current AI chatbots are generally capable of reproducing broad avulsion management frameworks. Most models consistently addressed general components such as replantation indications, antibiotic recommendations, tetanus assessment, and follow-up instructions. However, these elements represent general principles rather than scenario-

dependent clinical decisions. Similar observations have been reported in previous studies evaluating clinical decision-support tools and AI-assisted emergency triage systems.<sup>[10,11,21,22]</sup> Avulsion management is inherently algorithmic and unforgiving, and success depends not on general principles but on precise execution of time-dependent and scenario-specific steps, as emphasized in the IADT guidelines.<sup>[5,6]</sup> Recent studies evaluating large language models in dentistry have similarly demonstrated that AI systems may achieve acceptable overall performance while showing variability in clinically sensitive decision points requiring contextual interpretation.<sup>[15–17]</sup>

One of the key findings of this study was the presence of statistically significant between-model differences in root surface cleaning, splint type, and duration. These domains require a nuanced interpretation of the IADT guidelines and careful differentiation based on extra-oral dry time, storage conditions, and PDL viability.<sup>[3–5]</sup> These findings suggest that AI models may differ in their ability to integrate biological healing principles with procedural recommendations, rather than simply retrieving guideline information. This variability may reflect differences in how AI models prioritize procedural details versus general clinical principles, which has also been reported in previous evaluations of AI-based clinical decision-support tools.<sup>[15–17]</sup>

ChatGPT-5.2 and Perplexity AI consistently outperformed Gemini 2.5 Flash and DeepSeek-v3.2 in root surface cleaning recommendations. Correct differentiation between gentle saline rinsing, avoidance of mechanical scraping, and the use of fluoride treatment in delayed replantation scenarios is essential to reduce the risk of inflammatory or replacement resorption.<sup>[3,4]</sup> Inaccurate guidance at this step may irreversibly compromise periodontal healing, even if subsequent management steps are appropriate.

The most pronounced differences between chatbot models were observed in splint type and duration, which represent highly guideline-sensitive components of avulsion management. Flexible splinting for up to 2 weeks is a cornerstone of modern trauma care,<sup>[5]</sup> yet rigid splinting or inappropriate splint duration remains a common clinical error among practitioners.<sup>[7,8]</sup> The markedly lower accuracy observed for Gemini 2.5 Flash and DeepSeek-v3.2 in this criterion is clinically concerning, as incorrect splinting has been directly associated with increased rates of ankylosis and replacement resorption.<sup>[1,3]</sup> Splint selection represents a particularly sensitive decision point in avulsion management, as it requires integration of biological healing principles with mechanical stabilization strategies. Similar variability in AI performance has been reported in recent

dental AI studies, particularly in tasks requiring structured clinical judgment and protocol-based decision-making.<sup>[23–25]</sup>

Although some of the observed effect sizes were statistically moderate, their clinical implications may nevertheless be substantial. In avulsion management, even seemingly limited deviations from guideline recommendations in highly sensitive steps, such as splint type and duration, root surface management, or treatment decisions based on extra-oral dry time, can significantly influence PDL healing and pulpal prognosis. Unlike less critical supportive recommendations, errors in these domains may increase the risk of complications including inflammatory root resorption, replacement resorption, ankylosis, or long-term tooth loss. Therefore, moderate inter-model differences should not be interpreted solely as statistical variations, but rather in the context of their potential impact on clinical outcomes and evidence-based trauma management.

Although no statistically significant differences were observed in RHS values, the presence of high outliers indicates that even higher-performing models may occasionally generate inaccurate or fabricated references. Nevertheless, the presence of high outliers – particularly in ChatGPT-5.2 and Perplexity AI – underscores that even high-performing systems may occasionally generate severely inaccurate or fabricated citations. This finding is consistent with earlier reports documenting reference hallucination rates ranging from 30% to 60% in AI-generated medical and dental content.<sup>[12–14]</sup> These findings further emphasize that fluent and confident AI-generated outputs do not necessarily reflect factual accuracy, particularly in reference generation.

Given that the 2020 IADT guidelines serve as a definitive reference in dental traumatology, misquotation or fabrication of guideline citations poses a tangible risk, especially if AI-generated outputs are accepted uncritically in emergency settings. Even a single hallucinated or misattributed reference may mislead clinicians, educators, or trainees, potentially compromising evidence-based trauma management.<sup>[23,24]</sup>

From a clinical perspective, these findings suggest that AI chatbots should not be used as standalone decision-support tools for avulsion management. While they may assist in recalling general management principles, their variable performance in technically critical steps limits their reliability in real-time trauma care. Because the evaluations were based on adherence to IADT guidelines, responses categorized as inaccurate or partially compliant may also reflect recommendations that could be clinically inappropriate or potentially risky in real-life trauma management scenarios. Therefore, inter-model differences

should be interpreted not only as numerical performance variations but also in terms of their possible clinical consequences. This concern aligns with prior evaluations of AI use in emergency medicine and dentistry, which emphasize the continued necessity of clinician oversight and guideline literacy.<sup>[10,11,25–27]</sup> Therefore, AI-generated recommendations should be interpreted cautiously and verified against established clinical guidelines.

These findings also have implications for dental education and training. AI-generated responses that are partially correct may create a false sense of competence among students or junior clinicians, particularly when errors occur in less intuitive aspects of care such as splint selection or root surface management. Previous studies in dental education have highlighted the risk that AI tools may inadvertently reinforce superficial learning unless explicitly integrated into curricula emphasizing critical appraisal and guideline verification.<sup>[9,28–30]</sup> This highlights the importance of incorporating AI literacy into dental education curricula. Several limitations should be acknowledged. First, the use of a dichotomous (0/1) scoring system may oversimplify partially correct responses and does not capture nuances in clinical reasoning. In addition, all items were equally weighted, although certain criteria (e.g., splint type and duration, or management following prolonged dry time) may have greater prognostic significance. Second, the study evaluated a limited number of predefined scenarios, which, although carefully designed to represent key avulsion conditions, cannot encompass the full spectrum of clinical variability encountered in practice. Although the scenarios were developed according to the 2020 IADT guidelines, the inclusion of only 10 predefined scenarios may limit the robustness and generalizability of the findings. A larger and more diverse scenario set could reveal additional performance variability and potentially detect subtle inter-model differences that were not identified in the present study. Third, chatbot performance reflects a snapshot in time; ongoing model updates may alter compliance patterns. In addition, only a single response per scenario was evaluated, and potential variability across repeated queries was not assessed. Fourth, responses were assessed in English only, and performance may differ across languages. Fifth, the relatively small sample size and limited number of scenarios may have reduced the statistical power to detect subtle differences between chatbot models. Therefore, the findings should be interpreted with caution, and future studies with larger datasets are warranted. Finally, the study evaluated textual recommendations rather than real-world clinical behavior, which may be influenced by contextual and experiential factors not captured in prompt-based interactions.

Future research should expand scenario complexity, include additional trauma categories, and evaluate longitudinal changes in AI performance following model updates. Investigating clinician–AI interaction patterns and assessing whether AI-assisted decision-making improves or impairs adherence to trauma guidelines in simulated or real clinical environments would also be valuable, as suggested by previous work on digital decision-support systems in dental traumatology.

## Conclusion

Performance was higher in simpler scenarios, such as immediate replantation, whereas more complex conditions – particularly prolonged dry time and primary tooth avulsion – were associated with lower compliance and greater variability. Errors were most frequently observed in detailed clinical parameters, including splinting protocols and endodontic management. Reference accuracy also varied across scenarios, with higher hallucination scores observed in certain conditions, particularly in primary tooth cases, indicating inconsistencies in generating accurate bibliographic information. Overall, while AI chatbots demonstrate comparable overall performance in managing avulsion scenarios according to the 2020 IADT guidelines, notable discrepancies were observed in technically critical components that may directly influence clinical decision-making and prognosis. These findings suggest that aggregate performance scores may obscure clinically relevant differences in guideline-sensitive steps. Given that the analysis was based on a single response per scenario collected at a single time point, the results should be interpreted with caution, as chatbot outputs may vary across sessions and model updates. Accordingly, AI-generated guidance should be considered as a supportive educational resource rather than a substitute for established clinical protocols in dental trauma management.

**Ethics Committee Approval:** No human subjects or biological samples were involved; therefore, ethics approval was not required.

**Conflict of Interest:** The author declare that there is no conflict of interest.

**Financial Disclosure:** The authors declared that this study received no financial support.

**Use of AI for Writing Assistance:** The authors declared that artificial intelligence was not used in the study.

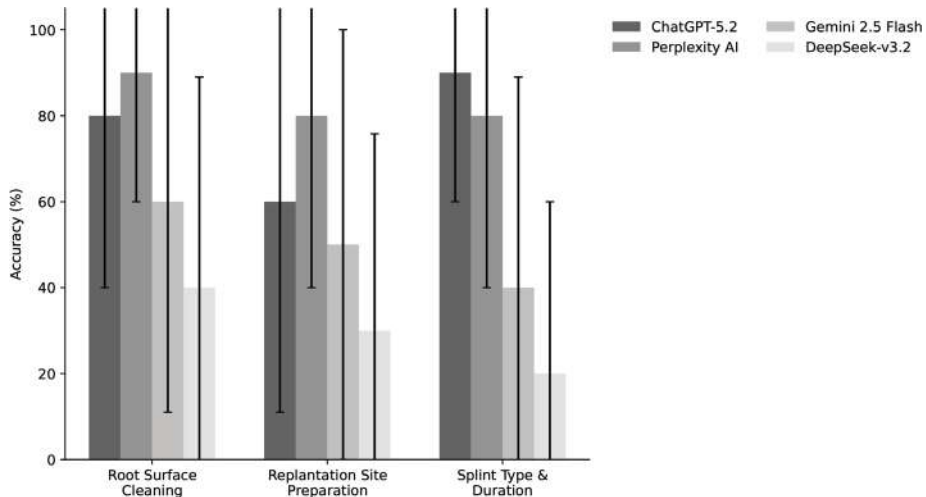
**Authorship Contributions:** Concept: EYM; Design: MO; Supervision: EYM, MO; Resource: EYM, MO; Data Collection and/or Processing: MO; Analysis and/or Interpretation: EYM; Literature Search: EYM, MO; Writing: EYM, MO; Critical Reviews: EYM, MO.

**Peer-review:** Double blind peer-reviewed.

## References

1. Bourguignon C, Cohenca N, Lauridsen E, Flores MT, O'Connell AC, Day PF, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 1. Fractures and luxations. *Dent Traumatol* 2020;36(4):314-30. [\[CrossRef\]](#)
2. Andreasen JO, Andreasen FM. *Essentials of traumatic injuries to the teeth: a step-by-step treatment guide*. 2<sup>nd</sup> ed. United States: Wiley-Blackwell; 2010.
3. Pohl Y, Filippi A, Kirschner H. Results after replantation of avulsed permanent teeth. II. Periodontal healing and the role of physiologic storage and antiresorptive-regenerative therapy. *Dent Traumatol* 2005;21(2):93-101. [\[CrossRef\]](#)
4. Trope M. Clinical management of the avulsed tooth: present strategies and future directions. *Dent Traumatol* 2002;18(1):1-11. [\[CrossRef\]](#)
5. Fouad AF, Abbott PV, Tsilingaridis G, Cohenca N, Lauridsen E, Bourguignon C, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 2. Avulsion of permanent teeth. *Dent Traumatol* 2020;36(4):331-42. [\[CrossRef\]](#)
6. Day PF, Flores MT, O'Connell AC, Abbott PV, Tsilingaridis G, Fouad AF, et al. International Association of Dental Traumatology guidelines for the management of traumatic dental injuries: 3. Injuries in the primary dentition. *Dent Traumatol* 2020;36(4):343-59. [\[CrossRef\]](#)
7. Al-Huthaifi BH, Ghwainem AA, Alqarni AS, Alshehri BY, Almnea RA, Alelyani AA, et al. Knowledge, perception, and management toward traumatic tooth avulsion among dental professionals: a cross-sectional study. *BMC Med Educ* 2025;25(1):1206. [\[CrossRef\]](#)
8. Mustuloğlu Ş, Deniz BP. Evaluation of Chatbots in the Emergency Management of Avulsion Injuries. *Dent Traumatol* 2025;41(4):437-44. [\[CrossRef\]](#)
9. Çege EE, Cömert H, Akal N, Ölmez A. Evaluation of the Performance of Artificial Intelligence Based Chatbots in Providing First Aid Information on Dental Trauma According to the ToothSOS Application. *Dent Traumatol* 2025;41(6):696-705. [\[CrossRef\]](#)
10. Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, et al. Evaluation of validity and reliability of AI Chatbots as public sources of information on dental trauma. *Dent Traumatol* 2025;41(2):187-93. [\[CrossRef\]](#)
11. Tokgöz Kaplan T, Cankar M. Evidence-Based Potential of Generative Artificial Intelligence Large Language Models on Dental Avulsion: ChatGPT Versus Gemini. *Dent Traumatol* 2025;41(2):178-86. [\[CrossRef\]](#)
12. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Med Inform* 2024;12:e54345. [\[CrossRef\]](#)
13. Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. *JAMA Ophthalmol* 2023;141(9):819-24. [\[CrossRef\]](#)
14. Sharun K, Banu SA, Pawde AM, Kumar R, Akash S, Dhama K, Pal A. ChatGPT and artificial hallucinations in stem cell research: assessing the accuracy of generated references - a preliminary study. *Ann Med Surg (Lond)* 2023;85(10):5275-8. [\[CrossRef\]](#)
15. Demir Cicek B, Cicek O. Evaluating the Response of AI-Based Large Language Models to Common Patient Concerns About Endodontic Root Canal Treatment: A Comparative Performance Analysis. *J Clin Med* 2025;14(21):7482. [\[CrossRef\]](#)
16. Yildirim A, Cicek O, Genç YS. Can AI-Based ChatGPT Models Accurately Analyze Hand-Wrist Radiographs? A Comparative Study. *Diagnostics (Basel)* 2025;15(12):1513. [\[CrossRef\]](#)
17. Yildirim A, Cicek O. Assessment of AI-Driven Large Language Models for Orthodontic Aesthetic Scoring Using the IOTN-AC. *Diagnostics (Basel)* 2025;15(23):3048. [\[CrossRef\]](#)
18. Guven Y, Ozdemir OT, Kavan MY. Performance of Artificial Intelligence Chatbots in Responding to Patient Queries Related to Traumatic Dental Injuries: A Comparative Study. *Dent Traumatol* 2025;41(3):338-47. [\[CrossRef\]](#)
19. Grinberg N, Arbel S, Boyadjiev YY, Ianculovici C, Kleinman S, Peleg O. The Performance of Artificial Intelligence in Providing Real-Time Aid in Emergency Dental Trauma: A Clinical Validation Study. *Dent Traumatol* 2026;42(3):356-62. [\[CrossRef\]](#)
20. Keleş ÖK, Arslan ZB. Performance of artificial intelligence chatbots in the diagnosis and management of simulated dental trauma cases: an evaluation based on IADT guidelines. *Clin Oral Investig* 2025;30(1):26. [\[CrossRef\]](#)
21. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial Intelligence and Primary Care Research: A Scoping Review. *Ann Fam Med* 2020;18(3):250-8. [\[CrossRef\]](#)
22. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv* 2023;55(12):1-38. [\[CrossRef\]](#)
23. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? *Am J Med* 2018;131(2):129-33. [\[CrossRef\]](#)
24. Borji A. A categorical archive of ChatGPT failures. *arXiv* 2023;2302.03494. [\[CrossRef\]](#)
25. Najeeb M, Islam S. Artificial intelligence (AI) in restorative dentistry: current trends and future prospects. *BMC Oral Health* 2025;25(1):592. [\[CrossRef\]](#)
26. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44-56. [\[CrossRef\]](#)
27. Lysaght T, Lim HY, Xafis V, Ngiam KY. AI-Assisted Decision-making in Healthcare: The Application of an

- Ethics Framework for Big Data in Health and Research. *Asian Bioeth Rev* 2019;11(3):299-314. [\[CrossRef\]](#)
28. Tiwari A, Kumar A, Jain S, Dhull KS, Sajjanar A, Puthenkandathil R, et al. Implications of ChatGPT in Public Health Dentistry: A Systematic Review. *Cureus* 2023;15(6):e40367. [\[CrossRef\]](#)
29. Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI. Impact of artificial intelligence on dental education: a review and guide for curriculum update. *Educ Sci (Basel)* 2023;13(2):150. [\[CrossRef\]](#)
30. Masters K. Artificial intelligence in medical education. *Med Teach* 2019;41(9):976-80. [\[CrossRef\]](#)



**Appendix Figure 1.** Comparison of chatbot performance across discriminative International Association of Dental Traumatology-based clinical criteria. Bars represent mean accuracy (%) and error bars indicate standard deviation. The figure displays performance across root surface cleaning, replantation site preparation, and splint type and duration for ChatGPT-5.2, Perplexity AI, Gemini 2.5 Flash, and DeepSeek-v3.2.

### Supplementary Material

- 1. Permanent maxillary incisor with closed apex – 30 min dry time**  
**Prompt:** A 12-year-old patient presents with an avulsed permanent maxillary incisor with a closed apex that has remained dry for 30 min. How should this case be managed according to the IADT Guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
- 2. Permanent maxillary incisor with open apex – 30 min dry time**  
**Prompt:** A 7-year-old child presents with an avulsed permanent maxillary incisor with an open apex that has remained dry for 30 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
- 3. Permanent maxillary incisor with closed apex – 90 min dry time**  
**Prompt:** A 12-year-old patient presents 90 min after avulsion of a permanent maxillary incisor with a closed apex, left completely dry. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
- 4. Permanent maxillary incisor with open apex – 90 min dry time**  
**Prompt:** A 7-year-old child presents with an open-apex permanent incisor avulsed and dry for 90 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
- 5. Permanent maxillary incisor with closed apex – stored in milk for 30 min**  
**Prompt:** A 12-year-old patient with a closed apex presents after an avulsion injury involving a permanent maxillary incisor, with the tooth stored in milk for 30 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
- 6. Permanent maxillary incisor with open apex – stored in milk for 30 min**  
**Prompt:** A 7-year-old child presents with an open-apex incisor that was stored in milk for 30 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
- 7. Permanent maxillary incisor with closed apex – stored in milk for 90 min**  
**Prompt:** A 12-year-old patient presents with an avulsed permanent maxillary incisor with a closed apex that was stored in milk for 90 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
- 8. Permanent maxillary incisor with open apex – stored in milk for 90 min**  
**Prompt:** A 7-year-old presents with an open-apex incisor stored in milk for 90 min. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
- 9. Avulsed primary maxillary incisor**  
**Prompt:** A young child presents with a completely avulsed primary maxillary incisor. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.
- 10. Immediately replanted tooth at the accident site (12-year-old patient)**  
**Prompt:** A 12-year-old patient presents with an avulsed permanent maxillary incisor with a closed apex that was immediately replanted at the accident site. How should this case be managed according to the IADT guidelines? Please also provide the full and accurate bibliographic reference for the IADT guideline on avulsion of permanent teeth, including the full title, all authors, journal name, year of publication, volume, issue, page numbers, and DOI.

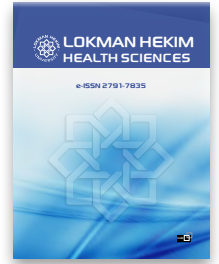
**Appendix Table 1.** Scenario-level IADT clinical compliance scoring matrix for all chatbot models across all evaluated avulsion scenarios

Scenario	Chatbots	Ifr	RSC	RSP	EC	AR	TP	STaD	Post-op	FUS	TS (0–9)
Closed apex – 30 min dry (12 years)	ChatGPT-5.2	1	1	0	1	1	1	0	0	1	6
Open apex – 30 min dry (7 years)	ChatGPT-5.2	1	0	1	1	1	1	1	1	1	8
Closed apex – 90 min dry (12 years)	ChatGPT-5.2	1	1	1	0	1	1	1	0	1	7
Open apex – 90 min dry (7 y)	ChatGPT-5.2	1	1	0	0	1	1	1	0	1	6
Closed apex – milk 30 min (12 years)	ChatGPT-5.2	1	1	0	1	1	0	1	0	1	6
Open apex – milk 30 min (7 years)	ChatGPT-5.2	1	1	0	1	0	1	1	0	0	5
Closed apex – milk 90 min (12 years)	ChatGPT-5.2	1	1	1	0	1	1	1	0	1	7
Open apex – milk 90 min (7 years)	ChatGPT-5.2	1	0	0	0	1	1	1	0	1	5
Avulsed primary maxillary incisor	ChatGPT-5.2	1	NA	1	NA	0	0	NA	1	0	3
Immediately replanted the tooth at the accident site (12 years)	ChatGPT-5.2	1	NA	NA	1	1	1	1	1	1	7
Closed apex – 30 min dry (12 years)	Perplexity AI	1	1	1	1	1	1	0	1	1	8
Open apex – 30 min dry (7 years)	Perplexity AI	1	1	1	0	0	1	0	1	1	6
Closed apex – 90 min dry (12 years)	Perplexity AI	1	1	1	0	1	1	0	1	0	6
Open apex – 90 min dry (7 years)	Perplexity AI	1	1	1	1	1	1	0	1	1	8
Closed apex – milk 30 min (12 years)	Perplexity AI	1	1	1	1	1	1	0	1	1	8
Open apex – milk 30 min (7 years)	Perplexity AI	1	1	1	1	0	0	0	0	0	4
Closed apex – milk 90 min (12 years)	Perplexity AI	1	0	1	1	1	1	0	0	1	6
Open apex – milk 90 min (7 years)	Perplexity AI	1	0	1	1	1	1	0	1	1	7
Avulsed primary maxillary incisor	Perplexity AI	1	NA	1	NA	0	0	NA	1	1	4
Immediately replanted tooth at the accident site (12 years)	Perplexity AI	1	NA	NA	0	1	1	0	1	1	5
Closed apex – 30 min dry (12 years)	Gemini 2.5 Flash	1	0	1	1	1	1	1	0	0	6
Open apex – 30 min dry (7 years)	Gemini 2.5 Flash	1	0	1	1	1	1	1	0	1	7
Closed apex – 90 min dry (12 years)	Gemini 2.5 Flash	1	0	1	1	1	1	1	0	0	6
Open apex – 90 min dry (7 years)	Gemini 2.5 Flash	1	0	1	0	1	1	0	0	1	5
Closed apex – milk 30 min (12 years)	Gemini 2.5 Flash	1	1	0	1	1	1	1	0	1	7
Open apex – milk 30 min (7 years)	Gemini 2.5 Flash	1	0	0	1	1	0	1	0	1	5
Closed apex – milk 90 min (12 years)	Gemini 2.5 Flash	1	1	1	1	1	1	1	1	1	9
Open apex – milk 90 min (7 years)	Gemini 2.5 Flash	1	1	1	1	1	1	1	1	1	9
Avulsed primary maxillary incisor	Gemini 2.5 Flash	0	NA	1	NA	0	1	NA	1	0	3
Immediately replanted tooth at the accident site (12 years)	Gemini 2.5 Flash	1	NA	NA	1	1	1	1	1	1	7
Closed apex – 30 min dry (12 years)	DeepSeek-v3.2	1	0	0	1	1	1	1	0	1	6
Open apex – 30 min dry (7 years)	DeepSeek-v3.2	1	0	0	1	1	0	1	0	1	5
Closed apex – 90 min dry (12 years)	DeepSeek-v3.2	1	0	0	0	1	1	0	1	1	5
Open apex – 90 min dry (7 years)	DeepSeek-v3.2	1	0	1	1	1	1	0	0	1	6
Closed apex – milk 30 min (12 years)	DeepSeek-v3.2	1	0	0	1	1	1	1	0	1	6
Open apex – milk 30 min (7 years)	DeepSeek-v3.2	1	0	1	1	1	1	1	0	1	7
Closed apex – milk 90 min (12 years)	DeepSeek-v3.2	1	0	1	1	1	1	0	1	0	6
Open apex – milk 90 min (7 years)	DeepSeek-v3.2	0	0	1	0	1	1	0	0	1	4
Avulsed primary maxillary incisor	DeepSeek-v3.2	1	NA	1	NA	0	0	NA	1	0	3
Immediately replanted tooth at the accident site (12 years)	DeepSeek-v3.2	1	NA	NA	0	1	1	1	1	1	6

IADT: International Association of Dental Traumatology; IFR: Indication for replantation; RSC: Root surface cleaning; RSP: Replantation site preparation; EC: Endodontic considerations; AR: Antibiotic recommendation; TP: Tetanus prophylaxis; STaD: Splint type and duration; Post-op: Post-operative care; FUS: Follow-up schedule; TS: Total score.

**Appendix Table 2.** Reference hallucination score for each chatbot-generated bibliographic reference across all scenarios

Scenario	ChatGPT-5.2	DeepSeek-v3.2	Gemini 2.5 Flash	Perplexity AI
Avulsed primary maxillary incisor	8	2	0	8
Closed apex – 30 min dry (12 years)	0	0	2	0
Closed apex – 90 min dry (12 years)	2	2	2	0
Closed apex – milk 30 min (12 years)	0	2	2	0
Closed apex – milk 90 min (12 years)	2	2	2	0
Immediately replanted tooth at the accident site (12 years)	2	2	0	2
Open apex – 30 min dry (7 years)	0	0	2	0
Open apex – 90 min dry (7 years)	0	2	2	1
Open apex – milk 30 min (7 years)	0	2	2	0
Open apex – milk 90 min (7 years)	2	2	2	0



# Publication Trends and Citation Patterns of Artificial Intelligence and Large Language Model Research in Orthopedics and Traumatology: A Türkiye-based Bibliometric Analysis

 Mehmet Can Gezer,  Mahircan Demir

Department of Orthopedics and Traumatology, Mamak State Hospital, Ankara, Türkiye

## Abstract

**Introduction:** Artificial intelligence (AI), particularly large language models (LLMs) and generative systems such as ChatGPT, has rapidly transformed medical research and clinical practice. Although global publication trends have been widely examined, country-specific bibliometric analyses in orthopedics and traumatology remain limited. This study aimed to perform a comprehensive bibliometric evaluation of Türkiye-based AI and LLM research in orthopedics and traumatology and compare the findings with global trends.

**Methods:** A cross-sectional bibliometric analysis was conducted in January 2026 using PubMed, TR Dizin, Scopus, and Web of Science databases. Original research articles from Türkiye-affiliated institutions involving AI, machine learning, deep learning, or large language model applications in orthopedics and musculoskeletal research were included. Journal Impact Factor (2024 JCR), citation counts, and SCImago Journal Rank (SJR) quartile classifications were recorded. Descriptive statistical analyses were performed.

**Results:** A total of 63 studies were included. Publication volume increased markedly after 2023. The mean Journal Impact Factor was  $2.34 \pm 1.15$  (median: 2.20; range: 0.8–5.4). Of the publications, 38.1% were in Q1 and 46.0% in Q2 journals, with none in Q4. The total citation count was 578, with a mean of  $9.17 \pm 14.20$  (median: 5; range: 0–82), reflecting a right-skewed distribution. Patient education studies were numerically predominant ( $n=34$ , 53.9%), while clinical application studies demonstrated the highest mean citation count ( $10.5 \pm 18.1$ ).

**Discussion and Conclusion:** AI-based research in orthopedics and traumatology in Türkiye has expanded rapidly since 2023, predominantly in Q1–Q2 journals. Citation patterns show considerable heterogeneity, consistent with an early developmental phase. These findings provide an objective bibliometric profile and may inform country-specific scientific strategies in AI-driven orthopedic research.

**Keywords:** Artificial intelligence; Bibliometrics; ChatGPT; Orthopedics; Surgery

**Cite this article as:** Gezer MC, Demir M. Publication trends and citation patterns of artificial intelligence and large language model research in orthopedics and traumatology: A Türkiye-based Bibliometric Analysis. Lokman Hekim Health Sci 2026;6(2):264–271.

**Correspondence:** Mahircan Demir, M.D. Mamak Devlet Hastanesi, Ortopedi ve Travmatoloji Kliniği, Ankara, Türkiye

**E-mail:** mahir19911991@gmail.com **Submitted:** 17.03.2026 **Revised:** 03.05.2026 **Accepted:** 25.05.2026 **Available Online:** 11.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



Artificial intelligence (AI) has evolved over the past decade from an experimental computational approach into a structural component of healthcare systems.<sup>[1,2]</sup> Machine learning, deep learning, and particularly large language models (LLMs) have generated a significant paradigm shift in modern medicine by enabling the analysis of high-dimensional clinical data, interpretation of imaging findings, comprehension of unstructured medical texts, and context-aware responses to complex clinical scenarios. Image-based diagnostic algorithms, risk prediction models, and natural language processing-driven systems have strengthened clinical decision-support mechanisms and accelerated the digital transformation of healthcare processes.<sup>[3,4]</sup>

With the widespread adoption of generative AI applications, transformation has extended beyond diagnostic and analytical domains to include scientific writing, academic productivity, patient education, and clinical decision-support systems.<sup>[5,6]</sup> This development indicates that AI is no longer merely an auxiliary tool but has become an active stakeholder in healthcare delivery and the research ecosystem.<sup>[7]</sup>

Orthopedics and traumatology are among the surgical disciplines in which AI applications have been rapidly integrated. Deep learning-based models have demonstrated promising levels of accuracy and efficiency in areas such as automated fracture classification, assessment of spinal deformities, arthroplasty planning, implant survival prediction, and interpretation of sports injuries. More recently, the evaluation of LLMs for generating responses to patient inquiries, assessing examination performance, providing clinical decision suggestions, and editing scientific manuscripts suggests that digital transformation in orthopedics has entered a new phase.<sup>[8–12]</sup>

However, the need for systematic evaluation of these technologies in terms of accuracy, methodological rigor, clinical applicability, ethical considerations, data security, and scientific impact has been increasingly emphasized in the literature. In particular, concerns regarding the potential for generative AI systems to produce “hallucinated” outputs, as well as issues related to source reliability and clinical validity, underscore the necessity for discipline-specific analyses.

In the international literature, AI-based publications in the field of musculoskeletal research have gained notable momentum, particularly in late 2022. Following the public release of generative AI platforms, studies evaluating

diagnostic accuracy, comparative clinical performance, quality of patient information, and educational reliability have increased rapidly.<sup>[12]</sup> However, existing bibliometric analyses indicate that AI-related research has been predominantly produced in North America, Western Europe, and East Asia, while contributions from developing countries have been analyzed only to a limited extent. Evaluating country-specific scientific positioning is important not only in terms of research output but also for assessing publication quality and international visibility.<sup>[13]</sup>

A noticeable increase in AI-related publications has also been observed in Türkiye-based orthopedics and traumatology research in recent years. Nevertheless, there is no comprehensive evaluation of the impact level of the journals in which these studies were published, their quartile distribution, citation performance, or their relative positioning within the global literature. In the current literature, Türkiye’s contribution in this field is largely interpreted through quantitative growth, whereas comparative analyses based on objective metrics of publication quality and scientific impact remain limited. This gap represents an important shortcoming in the development of country-specific strategic science policies.

Bibliometric analysis is a powerful methodological approach that enables the quantitative evaluation of scientific production within a specific thematic field through objective indicators.<sup>[12]</sup> Publication counts, citation metrics, journal impact indicators, quartile distributions, and temporal production trends provide an objective framework for assessing the scientific visibility and academic positioning of a research community. Particularly in rapidly evolving and interdisciplinary domains, bibliometric evaluations play a crucial role in identifying scientific orientation, quality trends, and potential areas for development.<sup>[13]</sup>

Although ChatGPT is one of the most prominent examples of generative AI, the present study was designed to include a broader spectrum of AI and LLM-based research to provide a more comprehensive bibliometric evaluation.

This study aimed to conduct a comprehensive bibliometric analysis of AI-based publications in orthopedics and traumatology affiliated with institutions in Türkiye. Accordingly, publication trends over time, journal quartile distributions (Q1–Q4), impact factor characteristics, and citation performance were evaluated. The findings were compared with general trends reported in the global literature in order to objectively determine Türkiye’s scientific positioning in AI-driven orthopedic research.

## Materials and Methods

This study was designed as a cross-sectional and descriptive bibliometric analysis to evaluate the bibliometric characteristics of AI-based publications in orthopedics and traumatology affiliated with institutions in Türkiye. This study was based on publicly available data and did not require ethical approval. The study selection process was reported using a PRISMA 2020 flow diagram (Fig. 1).

The literature search was performed in January 2026 using the PubMed, TR Dizin, Scopus, and Web of Science databases. For bibliometric verification, Journal Impact Factors (JIFs) and citation data were obtained from the Web of Science Core Collection, while quartile classifications were retrieved from the SCImago Journal Rank (SJR) database.

The search strategy was developed to capture studies related to AI, machine learning, deep learning, LLMs, generative AI, and ChatGPT within the fields of orthopedics, traumatology, and the musculoskeletal system.

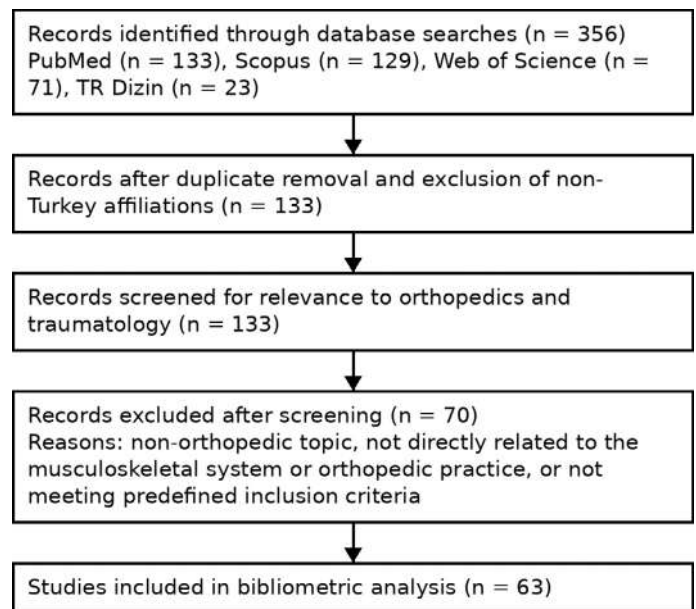
No country- or affiliation-based restriction was applied during the initial database search. Instead, eligibility regarding Türkiye affiliation was determined after record retrieval through manual screening of institutional addresses. Publications authored by Turkish researchers but not affiliated with institutions in Türkiye were excluded.

The search strategy was designed to identify studies related to AI, machine learning, deep learning, LLMs, generative AI, and ChatGPT within the fields of orthopedics, traumatology, and the musculoskeletal system.

Database-specific search strategies were adapted according to the indexing structure of each database. In TR Dizin, both Turkish and English keywords were used to enhance search sensitivity.

The complete search strategies for all databases are provided in Supplementary File 1.

Search results were filtered to include all available years without language restriction. No publication year restriction was applied during the search process. Due to substantial overlap across databases, duplicate records were identified and removed through both automated and manual screening. A total of 356 records were identified across all databases. After removal of duplicate records and exclusion of non-Türkiye affiliations, 133 records remained. Following screening, 63 studies were included. Following duplicate removal, records were screened based on title, abstract, and full text. In addition, institutional affiliations were evaluated to determine eligibility according to predefined inclusion criteria. The number of records



**Figure 1.** PRISMA 2020 flow diagram of the study selection process.

retrieved from each database and the complete search strategies for all databases are provided in Supplementary File 1 to ensure full reproducibility.

All identified records were independently screened at the title and abstract level by two researchers. Studies considered potentially eligible were subsequently assessed through full-text review. In cases of disagreement, consensus was achieved through discussion and, when necessary, consultation with a third reviewer.

Studies were included if they were conducted at institutions affiliated with Türkiye, were directly related to orthopedics, traumatology, or the musculoskeletal system, involved AI or related computational approaches, and were published as original research articles in peer-reviewed journals.

Exclusion criteria comprised letters to the editor, review articles, case reports, conference abstracts, studies unrelated to orthopedics, and publications not affiliated with institutions in Türkiye.

For each included study, the following variables were recorded: year of publication, journal name, country of publication, national or international classification of the journal, JIF (2024 JCR data), total citation count in the Web of Science, SJR quartile classification (Q1–Q4), and primary thematic category.

JIF values were obtained from the Journal Citation Reports (JCR), while quartile classifications were derived from the SJR database. These metrics were used complementarily to provide both impact- and ranking-based perspectives. Quartile classification was determined according to the

**Table 1.** Most cited studies included in the analysis

Study title	Citation count	IF	Quartile
ChatGPT performance in the medical specialty exam: An observational study	82	1.4	Q2
Use of Deep Learning Methods for Hand Fracture Detection from Plain Hand Radiographs	43	1.0	Q3
Large Language Models in Healthcare: A Bibliometric Analysis	45	2.4	Q2
Evaluating DeepResearch and DeepThink in Anterior Cruciate Ligament Surgery Patient Education: ChatGPT-4o Excels in Comprehensiveness, DeepSeek R1 Leads in Clarity and Readability of Orthopaedic Information	29	5.0	Q1
ViVGG19: Novel Exemplar Deep Feature Extraction-Based Shoulder Rotator Cuff Tear and Biceps Tendinosis Detection Using Magnetic Resonance Images	22	2.3	Q3
ChatGPT Can Offer At Least Satisfactory Responses to Common Patient Questions Regarding Hip Arthroscopy	22	5.4	Q1
Artificial Intelligence, Machine Learning, and Deep Learning in Orthopedic Surgery	21	1.0	Q2
Achieving High Accuracy in Meniscus Tear Detection Using Advanced Deep Learning Models with a Relatively Small Data Set	19	5.0	Q1
The Diagnosis of Developmental Dysplasia of the Hip From Hip Ultrasonography Images With Deep Learning Methods	16	1.5	Q2

IF: Impact factor; Q: SCImago journal rank.

journal's highest ranking within the orthopedics category or the most relevant subject category.

Data extraction was performed independently by two reviewers. Studies were categorized into three predefined thematic groups: patient-related applications, examination/education-related studies, and clinical applications. The classification process was conducted independently by two reviewers. In studies encompassing multiple thematic domains, categorization was determined based on the predominant content and the primary outcome of the study. Discrepancies between reviewers were resolved through structured discussion, and when necessary, consensus was achieved with the involvement of a third reviewer.

Studies employing various AI approaches (e.g., machine learning, deep learning, and LLMs) were included, and no subtype-specific categorization was performed, as the aim was to evaluate overall bibliometric characteristics.

None of the included studies was authored or co-authored by the present investigators.

### Statistical Analysis

The obtained data were analyzed using descriptive statistical methods. The distribution of continuous variables was assessed using both visual methods (histograms and Q-Q plots) and the Shapiro-Wilk test. Normally distributed variables were presented as mean  $\pm$  standard deviation (SD), whereas non-normally distributed variables were expressed as median (minimum-maximum) values.

Given the right-skewed nature of citation data, both mean and median values were reported to provide a comprehensive representation of central tendency. Categorical variables were presented as frequency (n) and percentage (%). Journal quartile distributions (Q1-Q4) were evaluated using frequency analysis.

Citation analysis included the total number of citations, mean and median citation counts per study, citation distribution according to quartile categories, and the proportion of uncited publications. The annual distribution of publications was illustrated graphically, and temporal trends were descriptively evaluated.

Due to the descriptive design of the study, no inferential statistical analyses were performed. All analyses were conducted using IBM SPSS Statistics for Windows, Version 26.0 (IBM Corp., Armonk, NY, USA).

### Results

A total of 356 records were identified through database searches, including PubMed (n=133), Scopus (n=129), Web of Science (n=71), and TR Dizin (n=23). Following the removal of duplicate records and studies not affiliated with institutions in Türkiye, 133 unique records remained for further evaluation. These records were screened for relevance to orthopedics and traumatology based on study titles, abstracts, and author affiliations. Studies not directly related to the musculoskeletal system or orthopedic practice were excluded. After eligibility assessment, 63 studies met the predefined inclusion criteria and were

**Table 2.** Bibliometric and Citation Analysis of Türkiye-Based Orthopedic AI and LLM-Based Studies (n=63)

	Total (n=63)	Patient education (n=34)	Exam/Board (n=9)	Clinical application (n=20)
Mean IF ( $\pm$ SD)	2.34 $\pm$ 1.15	2.41 $\pm$ 1.10	2.18 $\pm$ 1.20	2.33 $\pm$ 1.21
Median IF	2.20	2.30	2.00	2.20
IF (min–max)	0.8–5.4	0.9–5.4	0.8–4.7	0.8–5.4
Q1, n (%)	24 (38.1)	14 (41.2)	3 (33.3)	7 (35.0)
Q2, n (%)	29 (46.0)	15 (44.1)	4 (44.4)	10 (50.0)
Q3, n (%)	10 (15.9)	5 (14.7)	2 (22.2)	3 (15.0)
Q4, n (%)	0 (0)	0	0	0
Total citations	578	290	78	210
Mean citations ( $\pm$ SD)	9.17 $\pm$ 14.20	8.53 $\pm$ 12.5	8.66 $\pm$ 7.2	10.5 $\pm$ 18.1
Median citations	5	5	6	4
Citations (min–max)	0–82	0–60	0–21	0–82

IF: Impact factor; SD: Standard deviation; Q: SCImago Journal Rank; LLM: Large language models. Continuous variables are presented as mean $\pm$ SD; categorical variables as n (%).

included in the final analysis. The study selection process is summarized in Figure 1.

The distribution of publications by year is presented in Table 1. The number of publications by year was as follows: 2019 (n=1), 2020 (n=1), 2021 (n=2), 2022 (n=4), 2023 (n=10), 2024 (n=25), and 2025 (n=20). The distribution of publications by year is presented in Figure 2.

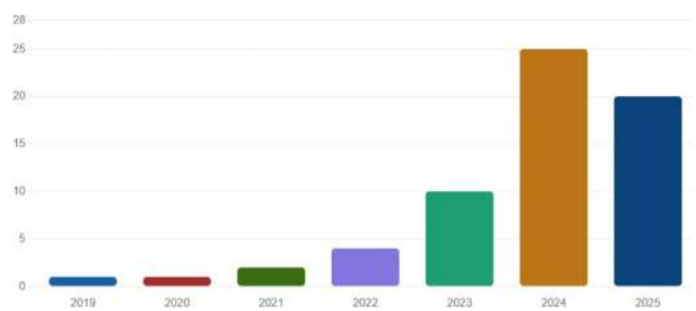
All included studies were published in international peer-reviewed journals. According to the 2024 JCRs data, the mean JIF was 2.34 $\pm$ 1.15, with a median of 2.20 (minimum: 0.8–maximum: 5.4).

The quartile distribution was as follows: Q1: 24 publications (38.1%), Q2: 29 publications (46.0%), and Q3: 10 publications (15.9%), while no studies were identified in Q4 journals. Publications in Q1–Q2 journals accounted for 84.1% of all studies (n=53).

The total number of citations was 578. The mean citation count per study was 9.17 $\pm$ 14.20, with a median of 5 (minimum: 0–maximum: 82). Eleven studies (17.5%) had not received any citations. The most cited studies included in the analysis are summarized in Table 1. Citation counts among the top-cited studies ranged from 16 to 82, encompassing publications related to specialty examinations, patient education, deep learning applications, and bibliometric analyses.

The distribution of studies across thematic subgroups was as follows: patient-related queries and patient education (n=34), specialty examination/board assessment (n=9), and clinical applications and diagnostic performance (n=20).

Bibliometric and citation characteristics according to subgroup classification are presented in Table 2.

**Figure 2.** Annual distribution of included publications (n=63).

## Discussion

In this study, the bibliometric profile of AI-based publications in orthopedics and traumatology affiliated with institutions in Türkiye was systematically analyzed. The findings indicate a noticeable increase in research output after 2023, with publications predominantly positioned in Q1–Q2 journals. The fact that 84.1% of the studies were published in Q1–Q2 categories may suggest that Türkiye-based orthopedic research has achieved a certain level of visibility and journal selectivity, rather than demonstrating a definitive competitive advantage. Large-scale bibliometric analyses of the global ChatGPT literature have reported Q1 publication rates generally ranging between 30% and 50%.<sup>[12,13]</sup> For example, a Web of Science-based analysis examining 2,465 articles reported a Q1 rate of approximately 40%.<sup>[12]</sup> Similarly, a more recent analysis of 3,231 articles demonstrated that publications were predominantly concentrated in Q2–Q3 categories, with journals indexed in ESCI representing a substantial proportion.<sup>[13]</sup> Early Scopus-based analyses have also reported relatively higher proportions of

Q3–Q4 publications.<sup>[14]</sup> The international comparisons presented in this study are based on studies using different databases, time frames, and inclusion criteria. Therefore, these comparisons do not represent direct statistical comparisons and should be interpreted with caution in light of methodological heterogeneity across studies.

In this context, the observed Q1 rate of 38.1% in our study appears to be within the range reported in the literature, rather than positioning Türkiye in a clearly defined global segment. Notably, no publications were identified in Q4 journals. International reports have indicated that a proportion of studies published in early 2023 appeared in lower-impact or rapidly reviewed journals.<sup>[14]</sup> The absence of Q4 publications among Türkiye-based orthopedic studies may reflect differences in publication patterns; however, this observation should be interpreted with caution, given the limited sample size and the early developmental stage of the field. The mean Impact Factor of  $2.34 \pm 1.15$  indicates that publications are clustered in journals with moderate impact. In the global ChatGPT literature, mean Impact Factor values have been reported to concentrate within the 2–3 range,<sup>[12,15]</sup> with slightly higher averages observed in analyses focused on healthcare disciplines.<sup>[15]</sup> Therefore, the Impact Factor profile of Türkiye-based orthopedic publications appears broadly comparable with the broader health sciences literature. In addition, recent studies evaluating the global bibliometric landscape of AI research in orthopedics have also reported a marked increase in publication output in recent years, with most studies being published in journals with moderate-to-high impact levels. The findings of the present study are generally consistent with global research trends in AI within the field of orthopedics.<sup>[16]</sup>

Studies focused on patient education were observed to have relatively higher mean Impact Factor values, paralleling global thematic trends. International bibliometric analyses have identified “medical education,” “ethics,” and “patient communication” as emerging motor themes in the generative AI literature.<sup>[13,17]</sup> Moreover, publications centered on education and patient communication have been reported to achieve early visibility and citation traction.<sup>[18–20]</sup> This pattern may indicate earlier adoption of generative AI in domains involving natural language generation and patient communication. The overall mean citation count of  $9.17 \pm 14.20$ , with a median of 5, reflects a right-skewed distribution. Given the right-skewed distribution of citation data, mean values may be influenced by a small number of highly cited studies. Therefore, median values were also reported to provide a more robust

representation of central tendency, and both measures were considered in the interpretation of the findings. Similarly, global analyses have reported high variance in citation distribution, with a limited number of studies accounting for a substantial proportion of citations.<sup>[12,15,21]</sup> Early-phase bibliometric evaluations have generally reported single-digit mean citation values.<sup>[14,22]</sup>

The presence of studies receiving up to 82 citations suggests that some early-published studies have reached relatively high citation levels within the field. Conversely, the finding that 17.5% of publications had not yet received citations is consistent with the phenomenon of “citation window bias.” Indeed, it has been reported that studies published after 2024 have, in the global literature as well, accumulated only limited citations to date.<sup>[13,23]</sup>

The higher mean citation rate observed for publications in Q1 journals is consistent with the commonly reported association between journal visibility and citation performance. Global bibliometric analyses have similarly demonstrated that ChatGPT-related publications in Q1 journals tend to accumulate citations more rapidly.<sup>[12,15,24]</sup> However, the observation of meaningful citation performance in Q2–Q3 journals suggests that multiple factors, including topical relevance, may contribute to citation patterns. In particular, healthcare-focused ChatGPT studies have been reported to achieve rapid citation uptake irrespective of JIF.<sup>[15,25]</sup> The thematic subgroup analysis indicates that research production in Türkiye initially evolved around patient education and information accuracy. This pattern aligns with global thematic network analyses, in which keyword co-occurrence networks have identified “education,” “ethics,” “accuracy,” “hallucination,” and “medical students” as dominant early-phase themes.<sup>[12,26]</sup> The relatively smaller number of studies focusing on clinical performance and diagnostic accuracy may reflect the current distribution of research topics within the dataset. Nevertheless, the higher mean citation value observed in this subgroup should be interpreted cautiously, as it may be influenced by variability and the presence of highly cited individual studies. International analyses have reported that clinically oriented or diagnostic accuracy studies demonstrate greater citation potential.<sup>[12,26]</sup>

The marked increase in publications after 2023 parallels the global production curve. However, the relative decline in the number of publications in 2025 should be interpreted with caution, as the literature search was conducted in January 2026 and some publications may not yet have been fully indexed across databases. In the international literature, 2023 has been characterized as the “year of exponential

growth” in ChatGPT-related publishing.<sup>[12,15]</sup> While the United States, China, and the United Kingdom have been reported as leading contributors in overall output, Türkiye’s simultaneous adaptation within a discipline-specific context appears to follow this global trend.<sup>[13,15]</sup>

This study has several important limitations. First, the relatively small number of included studies, particularly within certain subgroups, may affect the reliability of the descriptive statistics (mean, SD, and percentage distributions). Therefore, the presented findings should be interpreted with caution, and the generalizability of the results is limited. Subgroup-level analyses should be considered exploratory, and further studies with larger sample sizes are needed to validate these findings. Another important limitation of this study is the time-dependent nature of citation analysis. Since the included publications span different years, more recent studies have had a shorter period to accumulate citations. This may result in comparatively lower citation counts for studies published in 2025–2026. Citation counts were not normalized by publication year, which may further contribute to this bias. Therefore, direct comparisons of citation counts across publication years should be interpreted with caution. The JIF and quartile classifications used in this study are journal-level metrics and may not directly reflect the scientific quality of individual articles. This limitation has also been emphasized in the San Francisco Declaration on Research Assessment. Therefore, these metrics should be interpreted with caution when evaluating the findings. In this study, the bibliometric analysis was limited to descriptive statistics, and network-based analyses (such as co-authorship, co-citation, and keyword co-occurrence mapping) were not included. The primary aim of the study was to present the fundamental bibliometric characteristics of AI-based orthopedic research affiliated with Türkiye. Network-based bibliometric analyses (such as co-authorship, co-citation, and keyword co-occurrence mapping) may provide additional insights and could be considered in future studies.

## Conclusion

This bibliometric analysis demonstrates that AI-based research in orthopedics and traumatology in Türkiye has increased rapidly since 2023, with publications predominantly appearing in Q1–Q2 journals. Patient-related studies constitute the largest proportion of the literature, while citation counts show a heterogeneous distribution across studies. These findings reflect the current state of the field based on descriptive bibliometric indicators.

**Ethics Committee Approval:** This study was based on publicly available data and did not require ethical committee approval.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** During the preparation of this manuscript, AI-based writing tools were used for language editing and text improvement. The scientific content, data analysis, and conclusions are entirely the responsibility of the authors.

**Authorship Contributions:** Concept: MCG, MD; Design: MCG, MD; Supervision: MCG., MD; Data collection and/or processing: MCG; Analysis and/or interpretation: MD; Literature review: MCG; Writing: MD; Critical review: MCG.

**Peer-review:** Double blind peer-reviewed.

## References

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115-18. [\[CrossRef\]](#)
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44-56. [\[CrossRef\]](#)
3. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. *Annu Rev Biomed Eng* 2017;19:221-48. [\[CrossRef\]](#)
4. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A* 2018;115(45):11591-6. [\[CrossRef\]](#)
5. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med* 2018;15(11):e1002699. [\[CrossRef\]](#)
6. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop* 2018;89(4):468-73. [\[CrossRef\]](#)
7. Seah JCY, Tang CHM, Buchlak QD, Holt XG, Wardman JB, Aimoldin A, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Health* 2021;3(8):e496-e506. [\[CrossRef\]](#)
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2(2):e0000198. [\[CrossRef\]](#)
9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312. [\[CrossRef\]](#)
10. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating GPT as an adjunct for radiologic decision making: GPT-4

- versus GPT-3.5 in a breast imaging pilot. *J Am Coll Radiol* 2023;20(10):990-7. [\[CrossRef\]](#)
11. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst* 2023;47(1):33. [\[CrossRef\]](#)
  12. Nan D, Zhao X, Chen C, Sun S, Lee KR, Kim JH. Bibliometric analysis on ChatGPT research with CiteSpace. *Information* 2025;16(1):38. [\[CrossRef\]](#)
  13. Koo M. ChatGPT research: a bibliometric analysis based on the Web of Science from 2023 to June 2024. *Knowledge* 2025;5(1):4. [\[CrossRef\]](#)
  14. Farhat F, Silva ES, Hassani H, Madsen DØ, Sohail SS, Himeur Y, et al. The scholarly footprint of ChatGPT: a bibliometric analysis of the early outbreak phase. *Front Artif Intell* 2024;6:1270749. [\[CrossRef\]](#)
  15. Liu L, Qu S, Zhao H, Kong L, Xie Z, Jiang Z, Zou P. Global trends and hotspots of ChatGPT in medical research: a bibliometric and visualized study. *Front Med (Lausanne)* 2024;11:1406842. [\[CrossRef\]](#)
  16. Regmi A, Jain V, Baral S, Niraula BB. Global research landscape of artificial intelligence in orthopaedics: a decade of trends from a Scopus-based bibliometric analysis (2015-2025). *J Clin Orthop Trauma* 2026;73:103342. [\[CrossRef\]](#)
  17. Khan N, Khan Z, Koubaa A, Khan MK, Salleh RB. Global insights and the impact of generative AI-ChatGPT on multidisciplinary fields: a systematic review and bibliometric analysis. *Connect Sci* 2024;36:1-50. [\[CrossRef\]](#)
  18. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. *J Bus Res* 2021;133:285-96. [\[CrossRef\]](#)
  19. Aria M, Cuccurullo C. Bibliometrix: an R-tool for comprehensive science mapping analysis. *J Informetr* 2017;11(4):959-75. [\[CrossRef\]](#)
  20. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 2010;84(2):523-38. [\[CrossRef\]](#)
  21. Garfield E. The history and meaning of the journal impact factor. *JAMA* 2006;295(1):90-3. [\[CrossRef\]](#)
  22. Bornmann L, Daniel HD. What do we know about the h index? *J Am Soc Inf Sci Technol* 2007;58(9):1381-5. [\[CrossRef\]](#)
  23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71 [\[CrossRef\]](#)
  24. Huang X, Han F, Chen YF, et al. Bibliometric analysis of the application of artificial intelligence in orthopedic imaging. *Quant Imaging Med Surg* 2025;15(5):3993-4013. [\[CrossRef\]](#)
  25. Bernstein E, Ramsamooj A, Millar KL, Lum ZC. Identification and categorization of the top 100 articles and the future of large language models: thematic analysis using bibliometric analysis. *JMIR AI* 2025;4:e68603. [\[CrossRef\]](#)
  26. Zhang Y, Xie X, Xu Q. ChatGPT in medical education: bibliometric and visual analysis. *JMIR Med Educ* 2025;11:e72356. [\[CrossRef\]](#)

## Supplementary File

### Full Search Strategy and Database Records

#### 1. Databases

The literature search was conducted in January 2026 using PubMed, Web of Science Core Collection, Scopus, and TR Dizin.

#### 2. Search Strategies

PubMed:

("artificial intelligence"[Title/Abstract] OR "machine learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "large language model"[Title/Abstract] OR "generative AI"[Title/Abstract] OR "ChatGPT"[Title/Abstract]) AND ("orthopedic"[Title/Abstract] OR "orthopaedic"[Title/Abstract] OR "orthopedics"[Title/Abstract] OR "orthopaedics"[Title/Abstract] OR "traumatology"[Title/Abstract] OR "musculoskeletal"[Title/Abstract])

Web of Science:

TS=("artificial intelligence" OR "machine learning" OR "deep learning" OR "large language model" OR "generative AI" OR "ChatGPT") AND TS=("orthopedic" OR "orthopaedic" OR "orthopedics" OR "orthopaedics" OR "traumatology" OR "musculoskeletal")

Scopus:

TITLE-ABS("artificial intelligence" OR "machine learning" OR "deep learning" OR "large language model" OR "generative AI" OR "ChatGPT") AND TITLE-ABS("orthopedic" OR "orthopaedic" OR "orthopedics" OR "orthopaedics" OR "traumatology" OR "musculoskeletal")

TR Dizin: Turkish and English keywords were used.

#### 3. Record Counts

Database	Records Identified
PubMed	133
Web of Science	71
Scopus	129
TR Dizin	23

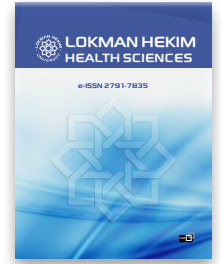
Total identified records across all databases: 133

#### 4. Data Processing

Duplicate records were removed using both manual and automated screening.

#### 5. Final Dataset

Final included studies: 63



# A Bibliometric Analysis of Artificial Intelligence and Simulation Technologies in Dental Education

 Hafize Gamze Demirbaş

Department of Restorative Dentistry, Aydın Adnan Menderes University, Faculty of Dentistry, Aydın, Türkiye

## Abstract

**Introduction:** Artificial intelligence (AI) and simulation technologies are increasingly used in dental education to improve learning safety, standardization, and skill assessment. This study examines their intersection, including AI-integrated simulations and curricula that apply AI and simulation separately. Despite rapid growth, the intellectual structure and trends of this intersection have not been comprehensively delineated in focused studies. We therefore conducted a bibliometric analysis of the literature.

**Methods:** Publications indexed in the Web of Science Core Collection from 2010 to 2026 were included. The search captured studies addressing dental education, simulation technologies, and AI applications. Bibliometrix in R was used to analyze publication trends, journals, countries, institutions, authors, and keyword patterns. Collaboration networks and keyword co-occurrence were visualized with VOSviewer. Only English-language articles and reviews were included.

**Results:** Publications increased markedly after 2019. Early research centered on virtual reality and simulation-based training, whereas recent studies emphasize AI, machine learning, and automated performance assessment. The United States and China were the most productive countries, with limited international collaboration. Keywords indicate a shift from immersive simulation toward AI-driven, adaptive, and data-informed educational systems. Both AI-integrated simulations and standalone AI or simulation applications were represented, reflecting diverse approaches.

**Discussion and Conclusion:** AI and simulation technologies constitute a growing research domain in dental education. The field is moving from technology-assisted simulation toward AI-enhanced systems, reflecting evolving research priorities rather than established clinical superiority. Future work should use standardized, multicenter, outcome-focused designs to assess educational effectiveness and relevance.

**Keywords:** Artificial intelligence; Bibliometric analysis; Dental education; Simulation; Virtual reality

Traditional apprenticeship models in dental education are increasingly complemented by safe, repeatable, and measurable technology-supported learning environments, reflecting a shift toward objective evaluation.<sup>[1]</sup> The COVID-19 pandemic accelerated hybrid

and technology-enhanced learning and highlighted the need for digital readiness in dental curricula.<sup>[2,3]</sup>

Simulation-based education enables translation of theoretical knowledge into clinical skills within controlled, risk-free environments. Immersive technologies such as

**Cite this article as:** Demirbaş HG. A Bibliometric Analysis of Artificial Intelligence and Simulation Technologies in Dental Education. Lokman Hekim Health Sci 2026;6(2):272–281.

**Correspondence:** Hafize Gamze Demirbaş, M.D. Aydın Adnan Menderes Üniversitesi Diş Hekimliği Fakültesi, Restoratif Diş Hekimliği Anabilim Dalı, Aydın, Türkiye

**E-mail:** hafize.demirbas@adu.edu.tr; gamzedemirbas91@gmail.com **Submitted:** 05.03.2026 **Revised:** 28.05.2026 **Accepted:** 08.06.2026

**Available Online:** 00.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



virtual reality (VR), augmented reality (AR), and haptic systems support interactive visualization, standardized practice, and objective performance assessment, strengthening psychomotor skill acquisition.<sup>[4-6]</sup>

Artificial intelligence (AI) has emerged as a transformative element in health professions education, supporting automated evaluation, adaptive learning pathways, and data-driven feedback. In dental education, AI applications have demonstrated benefits in diagnostic accuracy, learner confidence, and individualized formative assessment, particularly in clinical decision-making and radiographic interpretation.<sup>[7,8]</sup> Deep learning further augments cognitive and clinical training by identifying complex clinical patterns with high accuracy.<sup>[9]</sup>

The convergence of AI with simulation technologies advances intelligent learning environments. AI-supported haptic simulators enable objective assessment of manual performance, detection of technical errors, and immediate corrective feedback, supporting competency-based education and targeted remediation.<sup>[5,10-12]</sup> Implementation remains constrained by high costs, technical complexity, infrastructure requirements, and ethical concerns related to data privacy, algorithmic bias, and preservation of professional judgment.<sup>[11,13-15]</sup>

Despite the growing volume of studies, the intellectual structure and research dynamics at the intersection of AI and simulation technologies in dental education remain insufficiently specified rather than entirely unexplored. Prior bibliometric analyses show an overall increase in publications related to digital technologies in dental education, particularly following the COVID-19 pandemic.<sup>[16,17]</sup> While prior bibliometric studies have examined general trends in AI and digital intelligence within dentistry,<sup>[18-20]</sup> the novelty of this work lies in its integrated framework that maps AI-related approaches and simulation technologies both independently and in their convergence within dental education.

Therefore, this study aims to map the research landscape of AI and simulation technologies in dental education by conducting a bibliometric analysis of publications indexed in the Web of Science Core Collection between 2010 and 2026. By identifying publication trends, leading contributors, collaboration patterns, and emerging research themes, the study will clarify the field's developmental trajectory and provide a framework to guide future research, educational strategies, and policy development.

## Materials and Methods

### Study Design

This study was designed as a descriptive bibliometric analysis to examine the scientific literature on AI and simulation technologies in dental education. Since this analysis used only public bibliographic records, ethical approval was not required.

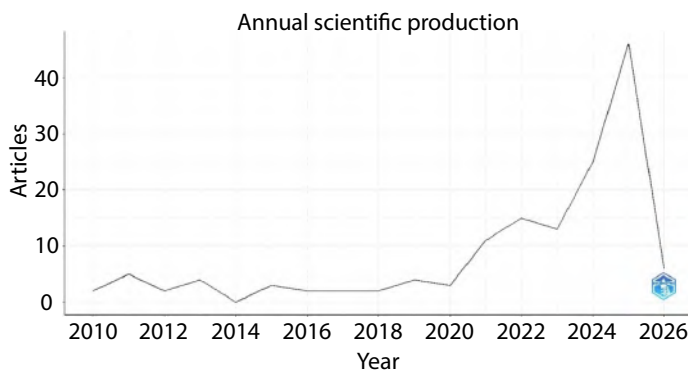
### Data Source and Search Strategy

The Web of Science Core Collection was selected as the primary data source due to its standardized citation indexing and suitability for bibliometric mapping in emerging scientific fields. To ensure methodological consistency and avoid potential duplicate indexing, a single standardized source was used throughout the analysis.

The database search was conducted on February 15, 2026, and covered publications indexed between January 01, 2010, and February 15, 2026. The year 2010 was selected as the starting point because it marks the beginning of a period in which digitalization, virtual simulation, and computer-assisted learning became increasingly visible in dental education, preceding the later acceleration of AI-based applications. Publications were retrieved using the Topic (TS) field, which includes titles, abstracts, author keywords, and Keywords Plus. The search strategy was designed to capture studies addressing dental education in conjunction with simulation technologies and AI-based applications. The following Boolean search query was applied:

```
TS=(dent*) AND TS=("dental education" OR training OR student* OR preclinical) AND TS=(simulation OR simulator OR "virtual reality" OR "augmented reality" OR haptic* OR "computer-based simulation") AND TS=("artificial intelligence" OR "machine learning" OR "deep learning" OR "intelligent system*" OR "computer-assisted" OR "automated feedback" OR "intelligent tutoring").
```

The search was restricted to articles and review papers published in English. Conference proceedings, editorial materials, letters, meeting abstracts, and book chapters were excluded to ensure methodological consistency and scientific rigor. To manage broad search terms such as "training" or "automated feedback," retrieved records were screened for topical relevance. Titles, abstracts, and keywords were reviewed to exclude clinical or engineering-focused studies unrelated to dental simulation or AI-supported assessment. During this screening stage, 14 records were excluded based on these criteria, reducing the initial pool of 159 publications to the final dataset of 145 publications. Broader e-learning records were retained



**Figure 1.** Annual scientific production related to artificial intelligence and simulation technologies in dental education (2010–2026).

only for contextual discussion rather than as direct evidence of effectiveness. All eligible records were exported in plain text format, including full bibliographic information and cited references.

## Data Analysis

Bibliometric analyses were conducted using the Bibliometrix package in R (Bibliometrix package version 5.1.1) in RStudio statistical software (version 2025.05.1+513; Posit Software, PBC, Boston, MA, USA) through its web-based interface, Biblioshiny. Descriptive performance indicators were calculated, including annual publication trends, leading authors, countries, institutions, journals, total citation counts, and h-index values.

To explore the intellectual structure and research dynamics of the field, established science mapping techniques were applied. These included analyses of co-authorship networks, keyword co-occurrence patterns, and citation relationships. Network visualizations were generated using VOSviewer software (version 1.6.20; Centre for Science and Technology Studies, Leiden University, Leiden, The Netherlands) to illustrate collaboration structures and thematic relationships. Threshold values for network inclusion were determined based on data distribution to ensure clear, interpretable, and reproducible visual representations. VOSviewer thresholds were selected to optimize readability and preserve the intellectual core by minimizing noise. A 3-occurrence minimum was applied to keyword co-occurrence to retain central themes while excluding idiosyncratic terms. Author-level networks used a 1-document and 1-citation minimum to reflect the field's emerging nature. These thresholds support reproducible construction and analytically meaningful cluster structures. No inferential statistical analyses were performed, as the study relied on bibliometric data derived from published literature.

From a methodological perspective, the bibliometric mapping procedures applied in this study followed established guidelines for conducting bibliometric analyses.<sup>[21,22]</sup> These approaches enable the identification of structural patterns, thematic evolution, and collaboration dynamics within a research field, providing a macro-level understanding of knowledge production and intellectual organization.

## Results

### General Characteristics of the Literature

The Web of Science search covering the period 2010–2026 initially identified a total of 159 publications; following the exclusion of 14 records based on the screening criteria, a final dataset of 145 publications focusing on AI and simulation technologies in dental education met the inclusion criteria and were included in the bibliometric analysis. These publications were distributed across 90 sources and comprised 106 original articles (including early access) and 39 review papers.

The dataset contained 6,037 cited references, with an average of 15.86 citations per document and a mean document age of 3.93 years. The annual publication growth rate was calculated as 7.11%, indicating a steady expansion of research activity in this field. Overall, 742 authors contributed; only 8 were single-authored, reflecting a collaborative structure. The mean number of co-authors per document was 5.34, and the international collaboration rate was 26.21%, suggesting a moderate level of global research cooperation.

In terms of thematic scope, the dataset included 577 author keywords and 303 Keywords Plus, indicating substantial conceptual and methodological diversity within the field.

The annual publication output remained relatively low and stable between 2010 and 2018, followed by a marked increase after 2019. Research productivity accelerated notably after 2021 and reached its highest level in 2025, while early indexing of publications in 2026 suggests continued growth momentum (Fig. 1).

Citation analysis showed that earlier publications achieved higher average citation counts, largely attributable to longer citation windows. The highest average citations per article were observed in 2013, while relatively strong citation performance was also noted for publications from 2021. Lower citation averages in the most recent years likely reflect limited time for citation accumulation rather than diminished scholarly impact.

**Table 1.** Top 10 most productive sources based on publication output and citation impact

Source	NP	TC	h-Index	First year
Journal of dental education	27	570	15	2010
European journal of dental education	5	43	3	2013
Journal of dentistry	4	44	4	2022
IEEE access	4	101	3	2024
Journal of medical internet research	3	112	3	2021
Applied sciences	3	73	3	2022
Cureus	3	47	2	2023
Dentistry journal	3	14	2	2025
Bioengineering	2	4	2	2023
International dental journal	2	7	2025	2

NP: Number of publications; TC: Total citations; h-index: Hirsch Index; First year: Year of the first publication of the source within the dataset.

Overall, research on AI and simulation technologies in dental education is a rapidly expanding domain with growing publications, collaboration, and scientific visibility.

### Most Productive Sources and Influential Documents

As noted in Section 3.1, the publications were disseminated across 90 sources. The journals represented in the dataset primarily covered dental education, health professions education, digital health, and computer-assisted learning, reflecting the convergence of educational sciences and emerging digital technologies.

The Journal of Dental Education was the most productive and influential source, publishing 27 articles and receiving 570 total citations (h-index=15). This was followed by the European Journal of Dental Education, which published five articles and accumulated 43 citations (Table 1). The

leading sources were mainly journals specializing in dental education and educational technologies, reflecting their central role in shaping research agendas and dissemination patterns. In addition, several interdisciplinary journals, such as those focusing on medical informatics, engineering, and applied sciences, also contributed to the literature, underscoring the integration of AI, VR, and simulation technologies into dental training.

Source impact analysis showed that journals with higher output demonstrated stronger citation performance, reflecting their influence in AI and simulation technologies in dental education. Publication concentration suggests knowledge production is anchored within core dental and interdisciplinary outlets (Table 1). Globally, the most cited studies addressed VR, haptic simulation, AI-based assessment, automated feedback, and adaptive learning environments. These publications reflect an increasing emphasis on objective evaluation, personalized learning, and immersive training in the field.

The most frequently cited document in the dataset was published by Frehywot et al.<sup>[23]</sup> receiving 338 total citations.<sup>[23]</sup> Although this article is broader in scope and focuses on e-learning in medical education rather than dental AI-simulation specifically, it appeared in the dataset because the search strategy captured broader digital education literature. It was therefore interpreted as a contextual citation reflecting digital education infrastructure and not as direct evidence for AI-supported dental simulation. Several more recent publications also exhibited high annual citation rates, indicating strong contemporary impact, particularly for studies integrating AI applications with digital and simulation-based learning approaches (Table 2).

**Table 2.** Top 10 most cited documents in the dataset

First author	Year	Source	Total citations	TC per year
Frehywot	2013	Human resources for health	338	24.14
Li	2021	Journal of medical internet research	99	16.50
Sai	2024	IEEE access	82	27.33
Schwendicke	2021	Journal of dental research	81	13.50
Ben Gal	2011	Journal of dental education	70	4.38
Sotgiu	2020	Anatomical sciences education	67	9.57
Monterubbianesi	2022	Applied sciences	58	11.60
Saghiri	2022	Journal of dental education	55	11.00
Collaço	2021	Journal of dental education	52	8.67
Gandedkar	2021	Seminars in orthodontics	42	7.00

TC indicates total global citations received according to the Web of Science database. TC per year represents the average annual citation rate.

These findings may suggest that research in AI and simulation in dental education is concentrated in a few influential journals and guided by highly cited studies on immersive simulation and AI-supported educational systems.

### Country and Institutional Contributions

A total of 45 countries contributed to the field, showing broad global engagement. China (n=147) and the United States (n=136) led scientific production, followed by India (n=41), South Korea (n=31), Germany (n=30), and Brazil (n=27). Significant contributions also came from Thailand (n=23), Canada (n=21), Japan (n=19), and Spain (n=18), spanning Asia, North America, and Europe.

Corresponding-author analysis (Table 3) showed that China and the USA each led with 26 publications (17.9%). However, collaboration profiles differed: China had a high single-country publication (SCP=92.3%), rate whereas the USA showed more international engagement (multiple-country publications [MCP]=30.8%). Germany and Thailand exhibited the highest relative international collaboration (MCP=66.7%), while Canada's output was entirely domestic.

At the institutional level, the output was concentrated in a few universities led by Sichuan University (n=28), the University of São Paulo (n=16), and Wuhan University (n=13). Other key contributors included Chulalongkorn, Danube Private, and Peking Universities (n=11 each), alongside Charité – Universitätsmedizin Berlin, the University of Iowa, and the University of Michigan. These findings highlight a landscape dominated by a few highly productive centers with uneven global collaboration.

### Author Productivity and Collaboration Network

A total of 742 authors contributed to the 145 publications included in the dataset, indicating a broad but highly dispersed authorship structure. The majority of authors produced only a single publication, suggesting that research activity in this field is distributed across many contributors rather than concentrated among a small group of consistently prolific researchers.

Only a limited number of authors published more than one article during the study period. The most productive contributors authored between two and three publications, with no individual exceeding three articles. Authors with the highest publication counts included Steven P. Engebretson, Peter Haddawy, Nikos Mattheos, Sallaberry, and Alice Urbankova, each contributing three publications. A second group of authors, including Richard Mosch, Siriwan Suebnukarn, A. Chaurasia, Xu Chen, and Yaning Li, each contributed two publications.

**Table 3.** Top 10 most productive countries based on corresponding author affiliations

Country	Articles	Articles (%)	SCP	MCP	MCP (%)
China	26	17.9	24	2	7.7
USA	26	17.9	18	8	30.8
India	9	6.2	7	2	22.2
Brazil	7	4.8	5	2	28.6
Germany	6	4.1	2	4	66.7
South Korea	6	4.1	4	2	33.3
Spain	6	4.1	4	2	33.3
Thailand	6	4.1	2	4	66.7
Saudi Arabia	5	3.4	3	2	40.0
Canada	4	2.8	4	0	0.0

SCP: Single-country publications; MCP: Multiple-country publications.

When co-authorship was taken into account using fractionalized authorship analysis, the proportional contributions of even the most productive authors remained limited. Fractionalized publication values ranged from approximately 0.24 to 1.17, reflecting substantial collaboration and the absence of sustained individual research dominance within the field.

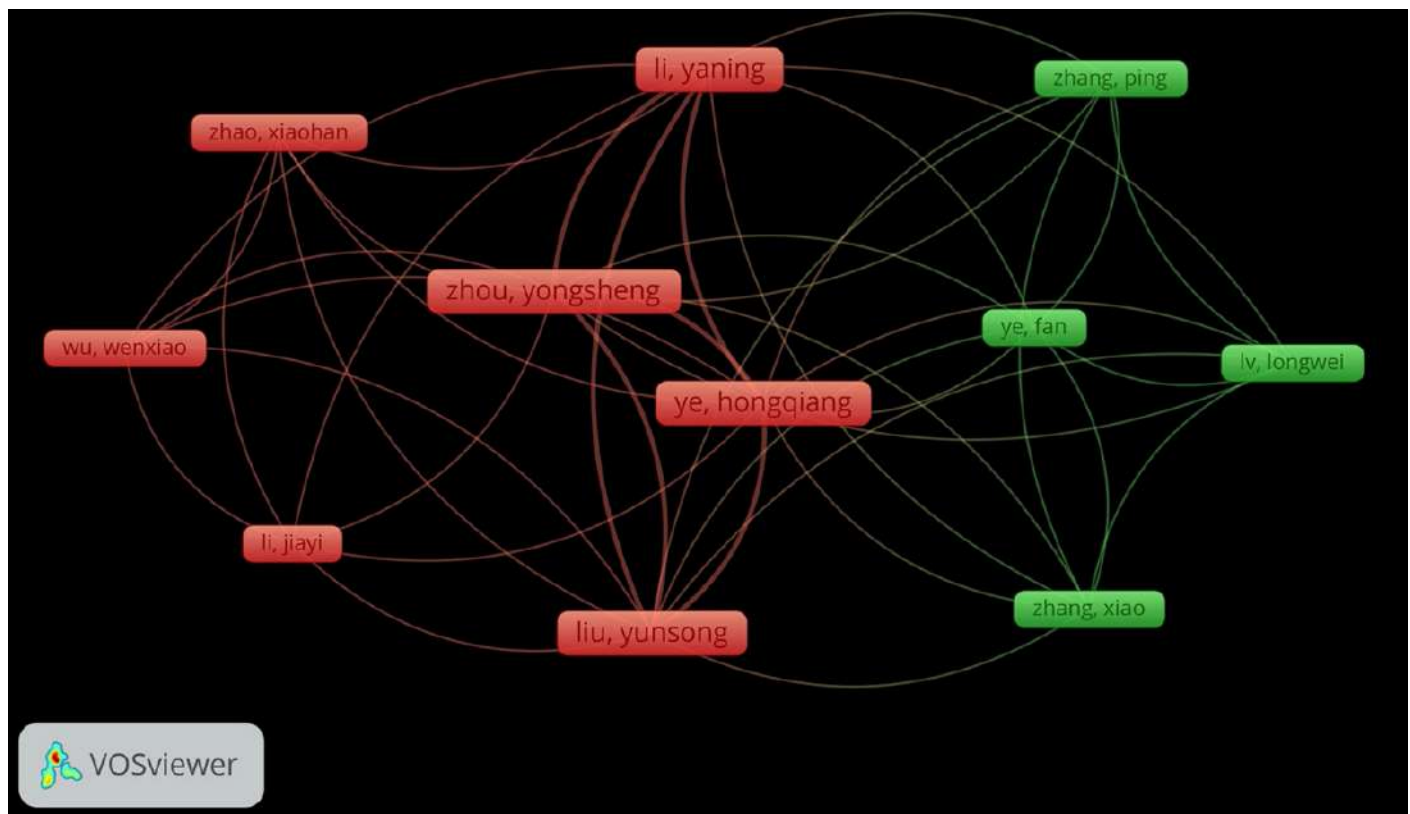
Author collaboration patterns were further examined through co-authorship network analysis using VOSviewer (Fig. 2). The resulting network revealed several small and weakly connected research clusters, with no evidence of large, densely interconnected collaboration groups. Most of the collaborative relationships were confined to small teams, and cross-cluster interaction was limited, indicating a fragmented collaboration structure at the author level.

To further characterize authorship patterns, Lotka's law was applied to assess the distribution of scientific productivity. The analysis demonstrated that 96.5% of authors contributed only a single publication, whereas only a very small proportion produced two or more papers. This highly skewed distribution is consistent with Lotka's inverse square law and confirms a high degree of author dispersion within the field.

Overall, the findings reveal limited author continuity and fragmented collaboration networks, suggesting that stable and consolidated research groups have not yet emerged in this domain.

### Keyword Analysis and Research Trends

A total of several hundred author keywords were identified across the 145 publications, reflecting the broad and multidisciplinary nature of research on AI and simulation



**Figure 2.** Author-level co-authorship network of publications on artificial intelligence and simulation technologies in dental education.

technologies in dental education. The most frequently occurring keywords were dental education ( $n=48$ ), AI ( $n=33$ ), dentistry ( $n=28$ ), simulation ( $n=27$ ), and VR ( $n=25$ ). Other commonly used terms included AR ( $n=20$ ), computer-assisted instruction ( $n=16$ ), performance ( $n=9$ ), students ( $n=9$ ), deep learning ( $n=8$ ), feedback ( $n=8$ ), and machine learning ( $n=5$ ), indicating a strong emphasis on both immersive technologies and data-driven educational approaches.

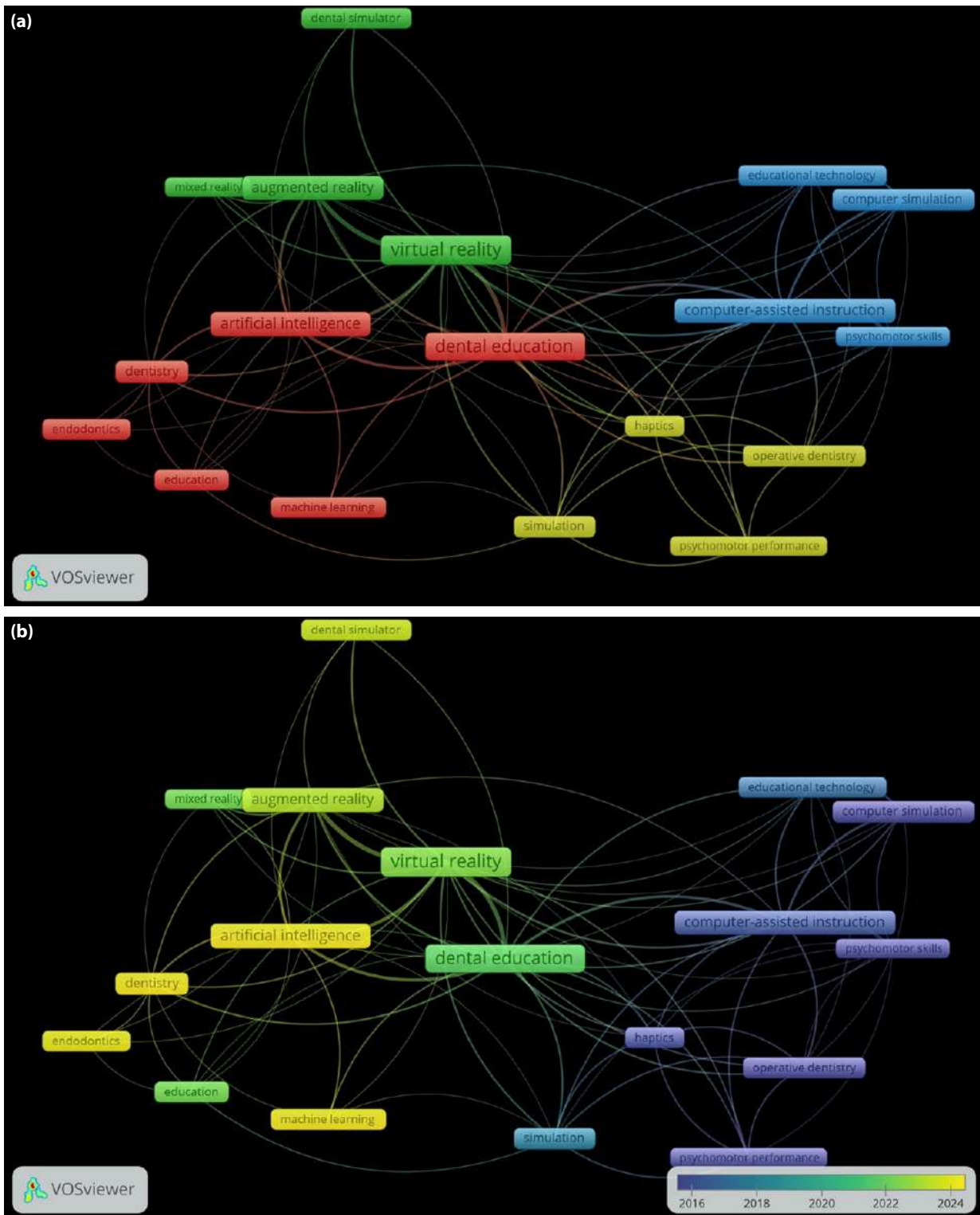
Keyword co-occurrence analysis conducted using VOSviewer revealed several major thematic clusters that together define the conceptual structure of the field (Fig. 3a). The largest cluster was centered on simulation-based dental education and included keywords such as simulation, VR, AR, computer simulation, simulator, and haptics. A second prominent cluster focused on AI-related concepts, including AI, machine learning, deep learning, model, and network, highlighting the increasing integration of intelligent systems into simulation-based training environments. A third thematic cluster represented educational processes and learning outcomes, with keywords such as students, learning, teaching, performance, assessment, and psychomotor skills. Smaller clusters reflected more specific clinical and technological

applications, including implant planning, digital dentistry, radiology, and computer-assisted surgical procedures.

Overlay visualization analysis demonstrated a clear temporal evolution in research focus (Fig. 3b). Earlier studies predominantly emphasized VR-based simulation systems and computer-assisted training tools. In contrast, more recent publications increasingly incorporated advanced AI concepts, including deep learning-based approaches and emerging discussions related to generative AI, alongside automated performance evaluation. This shift indicates a transition from technology-supported simulation toward intelligent, adaptive, and data-driven educational systems capable of personalized feedback and objective skill assessment.

## Discussion

This bibliometric analysis suggests that AI and simulation technologies are gaining increasing research attention in dental education, while also indicating that the field shows emerging but still uneven methodological development according to publication dispersion, collaboration structure, and keyword patterns. The pronounced increase in publication output after 2019 appears to coincide with pedagogical challenges highlighted during the COVID-19



**Figure 3.** Keyword co-occurrence network and temporal evolution of research topics in artificial intelligence and simulation technologies in dental education: **(a)** network visualization of author keyword relationships; **(b)** overlay visualization showing the temporal evolution of research topics.

pandemic and may reflect a broader transition in which digital technologies are increasingly discussed as integral components of dental education systems rather than temporary crisis-driven solutions.<sup>[2,16]</sup> These findings align

with systematic reviews indicating that recent literature increasingly frames digitalization not merely as an auxiliary enhancement but as a central component supporting standardization, measurability, and educational safety in

dental training.<sup>[1]</sup> These findings extend prior bibliometric studies on digital technologies and AI trends in dental education.<sup>[16–20]</sup> While earlier research mapped broader digitalization, this analysis provides a specific perspective by examining AI and simulation technologies within a shared dental education framework. By illustrating evolving thematic patterns and the interaction between these approaches, our study complements existing evidence and establishes a bibliometric foundation for future research in the field.

### **Bibliometric Structure and Global Research Dynamics**

Country- and institution-level analyses revealed a geographically concentrated research landscape, with China and the United States leading in overall publication output. However, corresponding-author-based indicators showed that countries such as the United States, Germany, and Thailand exhibit higher proportions of MCP, may suggest a more central role in shaping international research networks. In contrast, China's predominance of SCP appears to suggest strong domestic research capacity while indicating comparatively lower levels of international co-authorship within this dataset. This distinction appears consistent with prior bibliometric evidence that international collaboration enhances research visibility and knowledge diffusion.<sup>[16,17]</sup>

Author-level findings further support the interpretation that this research domain is still consolidating. The highly dispersed authorship structure, characterized by a large proportion of single-publication contributors, conforms to Lotka's law and may suggest that sustained, institutionally anchored research groups may still be in an early stage of development. Rather than indicating a lack of scientific interest, this pattern appears to reflect the interdisciplinary and exploratory nature of research on AI and simulation technologies, which often develops through short-term, technology-driven collaborations.<sup>[17]</sup>

### **Simulation Technologies: Educational Value Over Outcome Superiority**

Thematic and keyword analyses showed that early research predominantly focused on VR, haptic systems, and computer-assisted simulation platforms, whereas more recent studies increasingly emphasize AI-based assessment and feedback mechanisms. The literature has not reached a consensus on whether simulation technologies, with or without AI integration, produce superior clinical outcomes compared to conventional training. Instead, their primary contribution may lie in enhancing educational legitimacy through

standardization, repeatability, and process efficiency rather than demonstrable outcome dominance.<sup>[17]</sup>

In addition, some studies within the literature may suggest that highly interactive three-dimensional learning environments may increase cognitive load for certain learner profiles, particularly among students with lower spatial ability.<sup>[5,11,17]</sup> This observation appears to indicate that technological sophistication alone does not guarantee improved learning outcomes and that pedagogical alignment with learner characteristics remains critical.

### **AI and Automated Performance Assessment**

Bibliometric patterns appear to indicate that AI applications in dental education are most prominently concentrated in objective performance assessment and feedback systems. Machine learning- and deep learning-based approaches appear to be increasingly positioned as tools to reduce instructor dependency and enhance assessment consistency, functioning primarily as complementary mechanisms supporting formative assessment and competency-based education.

Beyond assessment, emerging work may suggest a gradual expansion of AI toward more integrative instructional roles, including intelligent tutoring and decision-support functions within simulation environments. This evolution can be interpreted in light of the concept of human-AI complementarity, although this interpretation should be treated as conceptual and linked cautiously to the observed emergence of keywords and studies related to intelligent tutoring, automated feedback, and decision-support functions.<sup>[24]</sup>

### **Ethical Considerations, Integration Challenges, and Future Directions**

Research on AI and simulation technologies in dental education appears to reflect ongoing technical and organizational challenges, including infrastructure demands and faculty preparedness, which may influence implementation and publication patterns.<sup>[11,15,25]</sup> Ethical considerations, such as data governance and algorithmic bias, have become more prominent in recent publications, as reflected by ethics-related keywords and thematic clusters, which may indicate an increasing emphasis on responsible innovation.<sup>[14]</sup> From a scholarly perspective, the observed fragmentation and thematic dispersion may suggest that future research may benefit from greater methodological rigor, multi-center collaboration, and outcome-oriented designs to strengthen the empirical foundation of AI and simulation technologies in dental education.

## Limitations and Drawbacks

This bibliometric analysis has several limitations. The study relied on a single bibliographic database, which may have excluded relevant publications indexed elsewhere. While this choice was made to improve the internal consistency of citation-based indicators, excluding databases such as Scopus, PubMed, and IEEE Xplore may limit coverage of interdisciplinary or conference-based studies not indexed in Web of Science. Variations in author names, institutional affiliations, or keywords could have influenced the results. Citation-based indicators reflect research productivity and visibility but do not directly assess methodological quality or educational effectiveness. Recent publications may appear less impactful due to time-dependent citation accrual, and early-access 2026 records may have incomplete citations. Finally, the interpretation of visualization maps and thematic clusters involves some subjectivity, so the findings should be considered indicative rather than definitive.

## Conclusion

This bibliometric analysis mapped the intersection of AI and simulation in dental education, identifying rapid growth since 2019. Research shows a shift toward AI-based automated assessment and intelligent tutoring within an increasingly connected international network. While integrated AI simulations are the primary focus, the bibliometric literature reviewed here does not yet reveal a consensus on whether AI-supported simulation produces superior clinical outcomes. Future research should adopt standardized, outcome-oriented designs to better evaluate educational effectiveness and clinical relevance.

**Ethics Committee Approval:** Ethics committee approval was not required for this study as it is a bibliometric analysis based on previously published literature and does not involve human participants or animal subjects.

**Informed Consent:** Informed consent was not required for this study as it is a bibliometric analysis based on previously published data and does not involve human participants.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

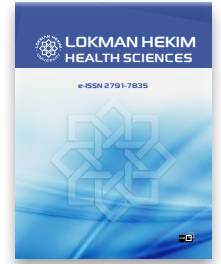
**Use of AI for Writing Assistance:** Generative AI tools were used solely to assist with language editing and the improvement of readability of the manuscript. No AI-assisted technologies were used in the study design, data collection, data analysis, or interpretation of results. All scientific content, analyses, and conclusions were generated, reviewed, and verified by the author, who takes full responsibility for the accuracy and integrity of the work.

**Peer-review:** Double blind peer-reviewed.

## References

1. Erdilek D, Gümüstas B, Efes BG. Digitalization era of dental education: A systematic review. *Dent Med Probl* 2023;60(3):513-525. [\[CrossRef\]](#)
2. Tadinada A, Gul G, Godwin L, Al Sakka Y, Crain G, Stanford CM, et al. Utilizing an organizational development framework as a road map for creating a technology-driven agile curriculum in predoctoral dental education. *J Dent Educ* 2023;87(3):394-400. [\[CrossRef\]](#)
3. Bashary NZ, Levine MH. Teaching strategy adaptations in undergraduate dental education during the COVID-19 pandemic. *J Dent Educ* 2024;88(6):865-71. [\[CrossRef\]](#)
4. Li Y, Ye H, Ye F, Liu Y, Lv L, Zhang P, et al. The current situation and future prospects of simulators in dental education. *J Med Internet Res* 2021;23(4):e23635. [\[CrossRef\]](#)
5. Huang S, Qu Y, Sun Y, Chen S, Gu L. Application and development of artificial intelligence and immersive technologies in dental education: a scoping review. *BMC Med Educ* 2025;26(1):89. [\[CrossRef\]](#)
6. Elhag HHM, Rao GKL, MohdNoor SNFB, Bin Zakaria MN, Mokhtar NB. Augmented reality in dental implants: A systematic review. *J Clin Diagn Res* 2024;18(1):20-5. [\[CrossRef\]](#)
7. Aminoshariae A, Nosrat A, Nagendrababu V, Dianat O, Mohammad-Rahimi H, O'Keefe AW, et al. Artificial intelligence in endodontic education. *J Endod* 2024;50(5):562-78. [\[CrossRef\]](#)
8. El-Hakim M, Anthonappa R, Fawzy A. Artificial intelligence in dental education: A scoping review of applications, challenges, and gaps. *Dent J* 2025;13(9):384. [\[CrossRef\]](#)
9. Cai J, Deng Y, Min Z, Zhang Y, Zhao Z, Jing D. Revealing the representative facial traits of different sagittal skeletal types: decipher what artificial intelligence can see by Grad-CAM. *J Dent* 2023;138:104701. [\[CrossRef\]](#)
10. Sallaberry LH, Tori R, Nunes FL. Automatic performance assessment in Virtual Reality medical simulators: A model based on procedure trajectories and machine learning. *Expert Syst Appl* 2024;238:122201. [\[CrossRef\]](#)
11. Lu B, Wei Z, Li X, Yin Y, Linghu J, Wang Y, et al. Progress of a novel dentistry teaching model based on the combination of virtual reality and artificial intelligence technologies in optimizing cognitive load: A systematic review. *J Dent Educ* 2025;0:1-18.
12. Vannaprathip N, Haddawy P, Schultheis H, Suebnukarn S. SDMentor: A virtual reality-based intelligent tutoring system for surgical decision making in dentistry. *Artif Intell Med* 2025;162:103092. [\[CrossRef\]](#)
13. Pimkhaokham A, Chow J, Pozzi A, Arunjaroen Suk S, Subbalekha K, Mattheos N. Computer-assisted and robotic implant surgery: Assessing the outcome measures of accuracy and educational implications. *Clin Oral Implants Res* 2024;35(8):939-53. [\[CrossRef\]](#)
14. Lin GSS, Foo JY, Goh SM, Alam MK. Exploring the ethical dimensions of artificial intelligence and robotics in dental education. *Bangladesh J Med Sci* 2024;23(4):999-1007. [\[CrossRef\]](#)

15. DaSilva AF, Robinson MA, Shi W, McCauley LK. The forefront of dentistry-promising tech-innovations and new treatments. *JDR Clin Trans Res* 2022;7(1\_suppl):165-245. [\[CrossRef\]](#)
16. Zou Z, Guo L. Insights into history and trends of teaching and learning in stomatology education: Bibliometric analysis. *JMIR Med Educ* 2025;11:e66322. [\[CrossRef\]](#)
17. Li Q, Li S, Fu D, Liao G, Zhou X, Gong T, et al. The role of emerging digital technologies in revolutionizing dental education: A bibliometric analysis. *J Dent Educ* 2025;90(5):753-63. [\[CrossRef\]](#)
18. Iniesta M, Pérez-Higueras JJ. Global trends in the use of artificial intelligence in dental education: A bibliometric analysis. *Eur J Dent Educ* 2026;30(2):427-38. [\[CrossRef\]](#)
19. Guo R, Zhang Y, Li W, Liu D, Hu W. Digital intelligence in dental education: A bibliometric analysis. *Int Dent J* 2026;76(3):109495. [\[CrossRef\]](#)
20. Long J, Yang L, Dou J, Zhang L, Tan X. Artificial intelligence in dentistry: A bibliometric analysis. *Br Dent J* 2025;1-7. [\[CrossRef\]](#)
21. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. *J Bus Res* 2021;133:285-96. [\[CrossRef\]](#)
22. Kumar R. Bibliometric analysis: comprehensive insights into tools, techniques, applications, and solutions for research excellence. *Spectr Eng Manag Sci* 2025;3(1):45-62. [\[CrossRef\]](#)
23. Frehywot S, Vovides Y, Talib Z, Mikhail N, Ross H, Wohltjen H, et al. E-learning in medical education in resource constrained low-and middle-income countries. *Healthc Hum Resour Health* 2013;11(1):4. [\[CrossRef\]](#)
24. Krumsvik RJ, Klock K, Bratteberg MH. Symbiotic intelligence in dental trauma diagnostics-an exploratory case study. *Front Oral Health* 2025;6:1687841. [\[CrossRef\]](#)
25. Uei LJ, Yeo XH, Leung YY, Pelekos G, Nawas BA, Mattheos N. Computer-Assisted Implant Surgery: Implications for teaching, learning, and educational strategies. *Clin Exp Dent Res* 2025;11(4):e70197. [\[CrossRef\]](#)



## ORIGINAL ARTICLE

# Evaluation of ChatGPT Responses to Questions About Joint Health and Physiotherapy in Patients with Hemophilia

Necati Muhammed Tat, Ayse Merve Tat

Department of Physiotherapy and Rehabilitation, Van Yüzüncü Yıl University Faculty of Health Sciences, Van, Türkiye

## Abstract

**Introduction:** This study aimed to evaluate whether ChatGPT provides satisfactory responses to frequently asked questions (FAQs) about joint health and physiotherapy approaches for patients with hemophilia.

**Methods:** Fifty questions were reviewed by five expert physiotherapists, who selected the 10 FAQs. Responses were generated using the GPT-4o model of ChatGPT on March 27, 2026. Two non-blinded reviewers independently evaluated the responses using a predefined four-point rating system based on accuracy, completeness, clarity, and clinical relevance. Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC) with a two-way random-effects model for absolute agreement. All responses were analyzed using the Flesch–Kincaid readability index to assess readability.

**Results:** The median response accuracy score (RAS) was 2, indicating that responses were generally satisfactory and required minimal clarification. Inter-rater reliability between the two reviewers was good (ICC=0.847). The Flesch–Kincaid score was 25, indicating that ChatGPT responses were difficult to read and understand by college graduates.

**Discussion and Conclusion:** ChatGPT demonstrated potential to improve basic knowledge of joint health in hemophilia, with excellent responses to 30% of FAQs. However, responses to physiotherapy-related questions often required additional clarification, particularly regarding exercise prescription parameters. Because physiotherapy interventions are individually tailored and require clinical supervision, ChatGPT may enhance patients' knowledge but cannot replace supervised physiotherapy practice. Therefore, ChatGPT should be considered a complementary tool rather than a primary clinical decision-making resource. Integrating artificial intelligence with physiotherapy expertise may enhance patient education and hemophilia care.

**Keywords:** Artificial intelligence; ChatGPT; Hemophilia; Joint health; Patient education; Physiotherapy

The use of artificial intelligence (AI) tools has become quite widespread in musculoskeletal rehabilitation, including physiotherapy.<sup>[1]</sup> Chat Generator Pre-Trained Transformer (ChatGPT) is a freely accessible program that attempts to provide responses to a wide range of questions from reinforcement learning using human feedback. ChatGPT

gained significant popularity, reaching over a billion monthly users within 4 months after its introduction in 2022. This widespread adoption of ChatGPT has significantly increased research publications investigating its potential utility in both academic and clinical settings. Continuous improvement in the AI tools over time is important for

**Cite this article as:** Tat NM, Tat AM. Evaluation of ChatGPT Responses to Questions About Joint Health and Physiotherapy in Patients with Hemophilia. Lokman Hekim Health Sci 2026;6(2):282–289.

**Correspondence:** Necati Muhammed Tat, PhD. Van Yüzüncü Yıl Üniversitesi, Sağlık Bilimleri Fakültesi, Fizyoterapi ve Rehabilitasyon Bölümü, Van, Türkiye  
**E-mail:** necatimuhammedtat@gmail.com **Submitted:** 21.04.2026 **Revised:** 09.06.2026 **Accepted:** 10.06.2026 **Available Online:** 16.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



**Table 1.** ChatGPT response rating system<sup>[9]</sup>

Accuracy description	Response accuracy score
1-Excellent response not requiring classification (The response was extremely accurate and comprehensive, providing information without the need for much explanation.)	1
2-Satisfactory requiring minimal clarification (The response was evaluated accurate; however, it required minimal additional clarification to address the user's question.)	2
3-Satisfactory requiring moderate clarification (The response is still accurate, but it is required a moderate level of further clarification to meet the user's needs).	3
4-Unsatisfactory requiring substantial clarification (The answer was deemed to be largely misinformed and overly generalized, which could cause misunderstanding by the user)	4

improving their reliability in physiotherapy practice. The reliability of physiotherapists' clinical reasoning using ChatGPT has also been demonstrated in the literature.<sup>[2]</sup> ChatGPT may have difficulties being readable for those below a university reading level.<sup>[3]</sup> ChatGPT has high potential for musculoskeletal rehabilitation education and debriefing but needs to be carefully integrated into practice due to its variability in complex clinical reasoning tasks.

Significant advances in pharmacological treatment have increased the life expectancy of patient with hemophilia (PwH) and allowed their active life. Besides pharmacological treatment options, non-pharmacological treatment modalities are critical for maintaining or improving joint health and physical function and for promoting regular physical activity.<sup>[4]</sup> Although physiotherapy has adapted to many changes in hemophilia in recent years, comprehensive care cannot be implemented in many physiotherapy centers, and there is great heterogeneity in practices.<sup>[5]</sup> With the COVID-19 pandemic, ensuring PwH has equal access to virtual counseling and tele-rehabilitation opportunities is critical to hemophilia care.<sup>[4]</sup> This situation can serve to help underserved PwH who have difficulty getting access to a physiotherapist with expertise in hemophilia to maintain and improve joint health.

A study focusing on the potential role of ChatGPT in total hip arthroplasty showed that it provided minimal to moderate explanations of frequently asked questions (FAQs) and could be used for patient education.<sup>[6]</sup> Another study reported that ChatGPT provided satisfactory to excellent responses to 60% of questions on ulnar collateral ligament reconstruction.<sup>[7]</sup> The use of ChatGPT in hemophilia is almost non-existent in the literature.

Joint health in hemophilia is a major cause of morbidity and continues to be a significant concern for PwH. During their life, PwH find themselves facing a wide variety of information and treatment choices regarding physiotherapy

approaches, manual therapy techniques, exercise type, physical activity recommendation, the duration and frequency of physiotherapy, and appropriate timing after a surgical operation. Using ChatGPT technology may have the possibility to both improve the knowledge of PwH regarding joint health and provide insights for clinical decision-making for physiotherapists who have no specialized expertise in hemophilia. The aim of this study is to determine whether ChatGPT provides satisfactory responses to FAQs about joint health and physiotherapy approaches in hemophilia. We hypothesized that ChatGPT would provide satisfactory responses to basic questions related to joint health and physiotherapy approaches in PwH.

## Materials and Methods

Five physiotherapists, who have expertise in hemophilia, were initially provided with 50 clinically relevant questions and instructed to generate a final list of 10 encountered FAQs in clinical practice, specifically related to joint health and physiotherapy (Appendix 1). The evaluators were not blinded to the model or question content during this selection process. Subsequently, and in accordance with a methodology previously described by Artioli et al.,<sup>[8]</sup> the resulting 10 FAQs were independently submitted to ChatGPT (GPT-4o), a large language model reported to be optimized for real-time responses. Queries were administered through the publicly accessible online interface on 27 March 2026.

Each response was recorded after the initial query without any follow-up prompts. All questions were presented in a standardized format to ensure consistency across inputs and to enhance the reproducibility of the assessment.

Response quality was assessed using an evidence-based four-point rating system adapted from Mika et al.<sup>[6]</sup> The system included four distinct categories (Table 1). Inter-rater reliability between the two reviewers was evaluated

**Table 2.** Evaluate ChatGPT's response accuracy score

Questions	RAS A1	RAS A2	Overall
Q1: What are the treatment options other than pharmacological treatment for joint health in hemophilia?	2	2	
Q2: What should I do in acute bleeding for joint health in hemophilia?	1	1	
Q3: Which physiotherapy approaches are used to improve joint health in hemophilia and can you prepare a sample treatment weekly plan?	3	2	
Q4: Which manual therapy techniques are used to improve joint health in hemophilia?	3	3	
Q5: What types of exercises are recommended to improve joint health in hemophilia?	2	2	
Q6: If I do not receive physiotherapy treatment for joint health, what changes will occur in my body in the long term?	1	1	
Q7: What are the potential risks of physiotherapy applications in hemophilic individuals?	1	1	
Q8: What should be the duration and frequency of physiotherapy for joint health in hemophilic individuals?	2	2	
Q9: Does fitness or physical activity improve joint health in hemophilia?	3	3	
Q10: What is the success rate of surgery for joint health problems in hemophilia and when should physiotherapy be started?	2	3	
Median accuracy score (interquartile range)	2 (1.5–2.5) 1.42–2.58	2 (1.5–2.5) 1.42–2.58	2 (1–3) 1.63–2.37

RAS: Response accuracy score.

using a two-way random-effects intraclass correlation coefficient (ICC) model with absolute agreement (ICC[2,1]) and reported with 95% confidence intervals (CI). Because RAS was ordinal (1–3 scale), descriptive statistics were summarized using median and interquartile range (IQR). ICC values were interpreted as follows: poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), and excellent (>0.90) reliability.

Readability was assessed using standardized readability formulas to evaluate how easy or difficult the responses generated by ChatGPT were to read and understand. The following indices were used: Flesch–Kincaid, Gunning Fog Index, Simple Measure of Gobbledygook, Automated Readability Index, and Dale–Chall Readability Score. This study did not require institutional review board approval.

## Results

### Response Accuracy Evaluation and Readability Checker

The list of FAQs asked to ChatGPT and the median response accuracy score (RAS) was 2 (%95 CI: 1.63–2.37), indicating that the answers were satisfactory and required minimal clarification in Table 2. Inter-class consistency (ICC) of ChatGPT responses was measured as (ICC=0.847). Responses to each question were pasted into a readability calculator and assessed using WordCalc software, where

the corresponding responses were recorded in Table 3. Readability scores indicate that the responses can be easily understood by the average college graduate.

### ChatGPT Response Analysis

ChatGPT responses to the 10 questions addressing joint health and physiotherapy management in PwH were presented in Appendix 1. Each response was independently evaluated by the reviewers for accuracy, completeness, and clinical adequacy.

#### **Question 1 Analysis: Satisfactory Response Requiring Minimal Clarification**

ChatGPT's response to non-pharmacological treatment options for joint health provides a comprehensive overview of the care in hemophilia. However, this answer does not provide information on which treatments to choose when considering the joint health status of PwH's. It also does not include the advantages and disadvantages of these treatment options compared to each other. It also does not provide any guidance on when to consider orthopedic intervention. Biopsychosocial multimodal approaches such as cognitive behavioral therapy, have been proven effective in reducing pain and kinesiophobia while improving quality of life and emotional status in PwH.<sup>[9]</sup>

**Table 3.** Individual readability level of ChatGPT's response

Questions	Flesch-Kincaid	Gunning Fog index	Simple Measure of Gobbledygook index	Automated Readability index	Dale-Chall
Q1	9.47	20.39	16.93	24.42	5.87
Q2	27.46	14.88	13.46	20.96	7.45
Q3	34.34	13.32	11.49	17.29	8.28
Q4	30.02	13.65	12.06	19.83	7.29
Q5	35.23	13.82	12.66	18.78	8.46
Q6	26.32	15.88	14.18	20.26	7.84
Q7	16.06	19.86	17.17	22.88	8.40
Q8	27.53	16.13	14.43	19.80	8.06
Q9	26.78	19.58	16.99	21.75	8.34
Q10	17.30	17.96	15.77	20.97	8.70
Average	25.05	16.54	14.51	20.69	7.86

### **Question 2 Analysis: Excellent Response Not Requiring Clarification**

ChatGPT wisely emphasizes that, just like the hemophilia guidelines, factor replacement should be done rapidly, and then bleeding should be controlled using the RICE principle.<sup>[10,11]</sup> The onset of pain, local discomfort, joint fullness, and a tingling sensation are the most common signs that bleeding has begun in the hemophilic individual.<sup>[10]</sup> If replacement therapy is not started immediately after the onset of joint bleeding, hemarthrosis will progress rapidly within a few hours, and the associated temporary functional impairment will develop into a painful, hot, and swollen joint. Administration of clotting factors rapidly relieves pain, but inflammation and functional impairment may persist for some time, depending on the amount of blood in the joint. The information given regarding the start of physiotherapy is particularly accurate. Prolonged rest periods may cause muscle weakness and loss of joint control, while early mobilization carries the risk of re-bleeding. Physiotherapy should be started once the acute symptoms of swelling, redness, and loss of movement have subsided.<sup>[12]</sup> A gradual strengthening program is also the right approach afterwards.

### **Question 3 Analysis: Satisfactory Response Requiring Moderate Clarification**

The most commonly used methods were physiotherapeutic approaches, such as strengthening exercises, education, range of motion exercises, soft tissue techniques, taping, and manual therapy, while less common methods were massage, electrotherapy, splinting, and ultrasound therapy.<sup>[5]</sup> However, the response does not address kinesiotherapy, electrotherapy, ultrasound therapy, and massage. The

recommendation to work closely with a hematologist and physiotherapist specializing in hemophilia care is advised for PwH to assess the factor level in each patient and to select physiotherapy approaches based on the joint health status life-span of life.<sup>[13]</sup>

### **Question 4 Analysis: Satisfactory Response Requiring Moderate Clarification**

Manual techniques are given subheadings and in general terms. For example, "How many strokes were made in myofascial techniques, how many deep and superficial techniques were applied?"<sup>[14]</sup> "Which technique was used in joint mobilizations?"<sup>[15]</sup> "Was a traction pattern applied in joint mobilizations?"<sup>[15]</sup> How the movements were made should be stated. The techniques used, the methodology, and their characteristics are not explained in detail.

### **Question 5 Analysis: Satisfactory Response Requiring Minimal Clarification**

In the Chat GPT response, a range of motion, strengthening, proprioception, balance, stretching, functional, and low-impact aerobic exercises were explained in detail with examples. In the systematic review evaluating the effect of exercise on postural balance in hemophilia, it is seen that these exercises are given as an intervention.<sup>[16]</sup> However, the questions "How long were these exercises done? Were there any pauses between sets/repetitions? How many series were developed?" were not answered. In the systematic review reporting exercise variables and pain threshold for strength training protocols in PwH, variables such as frequency, duration, intensity of exercise, whether the exercise was supervised, whether any equipment was needed, and exercise progression were evaluated.<sup>[17]</sup>

### **Question 6 Analysis: Excellent Response not Requiring Clarification**

The possible conditions and effects that may occur if they do not receive physiotherapy treatment are clearly and accurately stated, just as stated in the hemophilia guidelines. Recurrent joint bleeding in hemophilia may lead to irreversible damage to the musculoskeletal system as a late symptom.<sup>[10]</sup> Pain and depression/anxiety correlate with the severity of hemophilia and have negative effects on health-related quality of life.<sup>[18]</sup> It has been reported that PwH may be able to preserve joint health with improved treatment strategies. Surgical interventions for hemophilia can be performed at a younger age and may carry a higher risk of complications,<sup>[19]</sup> but with physiotherapy treatment strategies, joint health can be improved and surgical procedures can be postponed to older ages.

### **Question 7 Analysis: Excellent Response Not Requiring Clarification**

Physiotherapy must be approached cautiously due to the unique risks involved. Improper or aggressive physiotherapy can lead to complications due to the bleeding tendency inherent in hemophilia. The fact that ChatGPT particularly touched upon the issues of lack of qualified supervision, the importance of biomechanics, and exercise timing as risks of physiotherapy are signs that evaluated the subject holistically. Individualized exercise program, slow and gradual progression, factor replacement before treatment, and multidisciplinary approaches recommendations are very important to reduce risks. It has been shown in the literature that the supervised physiotherapy program was successful in improving joint health, functional capacity, quality of life, risk of fall, and kinesiophobia.<sup>[20]</sup>

### **Question 8 Analysis: Satisfactory Response Requiring Minimal Clarification**

ChatBot correctly highlights that the frequency of physiotherapy must be carefully tailored to their specific needs and current condition. It also states that the frequency of physiotherapy will be specifically adjusted in supervised programs, home exercise programs, and periods after bleeding or surgery. However, no response was provided on the duration of physiotherapy sessions and exercise duration. Through interaction with children with hemophilia and their families, the DOLPHIN Exercise program was designed and developed to include weekly stretching, balance, strengthening, and functional activities.<sup>[21]</sup>

### **Question 9 Analysis: Satisfactory Response Requiring Moderate Clarification**

ChatGPT explained in a comprehensive overview of the effects of fitness and physical activity on the body of a hemophiliac individual, including precautions. However, the response does not address the use of factor before fitness or physical activity or the minimum factor level. It did not provide sufficient information to avoid high-impact activities such as running and soccer.

In today's age of new treatments, hemophilic individuals need to be encouraged to engage in physical activity. In the study evaluating the types of physical activity of 374 hemophilic individuals, their participation in low-impact activities such as walking, swimming, or cycling was high (90.9%).<sup>[22]</sup> Nowadays, PwH's participate in competitive sports with factor therapy. Fitness was found to improve not only joint health but also psychosocial well-being and self-esteem in young men with hemophilia aged 18–25 years, with participants also reporting additional benefits such as less kinesiophobia, reduced pain, weight loss, and improved self-esteem.<sup>[23]</sup> Low physical activity levels in PwH may significantly affect the risk and severity of abdominal fat and hypertension and various cardiometabolic and/or musculoskeletal health problems in aging hemophiliacs.<sup>[24]</sup> Therefore, regular multimodal training is important for PwH, and the first priority for PwH to participate in this type of program is to be under adequate factor treatment. To provide safety, fitness or physical activity individually adapted to the PwH's needs should be prescribed and supervised by a physiotherapist experienced in the field of hemophilia.<sup>[23]</sup> It has been shown that the participation of PwH in physical activity was affected not only by their own abilities and attitudes but also by family, friends, and health professionals.<sup>[25]</sup>

### **Question 10 Analysis: Satisfactory Response Requiring Minimal Clarification**

ChatGPT accurately states that the time to start physiotherapy should be as soon as possible but it is specific to the type of surgery, the severity of hemophilia, and the individual's status. However, no information was provided in the response regarding the exercise content in the post-operative physiotherapy phases. The use of clotting factor replacement allows earlier and more intensive rehabilitation in the post-operative period in PwH. Various orthopedic interventions can be applied to PwH for joint health: Arthroscopic synovectomy and joint debridement, radial head resection, tibial

osteotomy, arthrodesis, tendon lengthening, and total joint arthroplasty. In the hemophiliac and factor VIII inhibitor case undergoing total knee replacement, supervised physiotherapy has been shown to benefit from combining 6 weeks of pre-operative physiotherapy with 6 weeks of post-operative rehabilitation to increase the range of motion and muscle strength, facilitate rapid mobilization and recovery of function, and minimize the risk of bleeding.<sup>[26]</sup> Although the pre-operative range of motion (ROM) is the most important variable affecting postoperative ROM in total knee arthroplasty, early mobilization should be offered in the 1<sup>st</sup> post-operative days to improve the outcome. Isometric contractions and gentle open kinetic chain exercises should also be initiated to prevent quadriceps inhibition and restore control. In recent years, structured exercises consisting of closed kinetic chain exercises have been shown to produce improvements in physical activity, balance, proprioception, and gait in PwH's.<sup>[27]</sup> As part of comprehensive care, both pre- and post-operative physiotherapy is essential to achieve optimal functional outcome and thus optimal quality of life in PwH.

## Discussion

This study evaluated the quality of ChatGPT-generated responses related to joint health and physiotherapy management in PwH. Overall, ChatGPT demonstrated acceptable performance, with 30% of responses rated as excellent, 40% requiring minimal clarification, and 30% requiring moderate clarification. The descriptive comparison across question categories suggests that ChatGPT's responses have a more consistent performance in general educational content, and clinically detailed questions required greater clarification. These findings suggest that ChatGPT may support health literacy and acquisition, particularly among patients with limited access to specialized physiotherapy services, caregivers, and non-specialist clinicians.

ChatGPT demonstrated high accuracy (80%) and strong reliability in aligning its responses with evidence-based information when used as a clinical decision-support tool in physical therapy.<sup>[1]</sup> Similarly, satisfactory responses (90%) have been reported for preoperative questions in total hip arthroplasty,<sup>[6]</sup> while lower rates (60%) were observed in UCL reconstruction.<sup>[7]</sup> McCarthy et al.<sup>[28]</sup> reported that ChatGPT-generated accurate responses in cases (88%), although patients perceived professionally developed educational materials as more useful. In patellar tendinopathy, ChatGPT provided comprehensive information but exhibited

limitations in accuracy and readability.<sup>[3]</sup> Consistent with these findings, ChatGPT has also been identified as a reliable and accessible source of information for caregivers of PwH, particularly for FAQs, while remaining a complementary resource rather than a substitute for healthcare professionals.<sup>[29]</sup> Given the individualized nature of physiotherapy interventions and exercise prescription, clinical decision-making should continue to rely on professional assessment and supervision. Although future advances in artificial intelligence may facilitate more personalized rehabilitation recommendations,<sup>[30]</sup> AI-based tools are currently best viewed as educational and decision-support resources that can enhance patient knowledge rather than replace physiotherapy practice.

Although ChatGPT demonstrated relatively high Flesch–Kincaid readability scores and may be useful for patient education, readability alone does not reflect clinical accuracy, safety, or contextual appropriateness. Consequently, AI-generated content should not be used independently by patients, particularly in complex conditions such as hemophilia that require individualized hemophilia care. Instead, ChatGPT should be viewed as an adjunctive educational tool that can assist healthcare professionals in generating patient-friendly explanations, while clinical oversight and content validation remain essential.<sup>[3]</sup>

## Conclusion

Despite these potential benefits, several important limitations should be acknowledged. Large language models may generate inaccurate or fabricated outputs, which may pose risks in hemophilia, particularly in relation to bleeding management and exercise prescription. Readability metrics, while useful for assessing linguistic accessibility, do not capture clinical validity, safety, or contextual appropriateness. In addition, model updates, prompt sensitivity, and inherent system variability may affect response consistency, thereby limiting reproducibility and cross-study comparability. An additional limitation of this study is the potential for evaluator bias. Although response quality was assessed using predefined criteria and expert review, the evaluation process inevitably involved subjective judgment, which may have influenced the assigned ratings. Furthermore, the rating framework used in this study, although widely applied in the literature, lacks external validation. Therefore, AI-generated outputs should be interpreted cautiously and verified against evidence-based guidelines and clinical expertise before being incorporated into patient care.

**Ethics Committee Approval:** Ethical approval and informed consent were not obtained as publicly accessible Chat Generator Pre-Trained Transformer (ChatGPT) is a free and openly accessible program were used and no human/animal participants were included in the study. This study did not require institutional review board approval.

**Informed Consent:** Informed consent were not obtained and human participants were not included in the study.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** ChatGPT (GPT-4o) has been used.

**Authorship Contributions:** Concept: NMT; Design: NMT; Supervision: AMT; Data collection and/or processing: NMT, AMT; Analysis and/or interpretation: NMT, AMT; Literature review: NMT, AMT; Writing: NMT; Critical review: AMT.

**Peer-review:** Double blind peer-reviewed.

## References

- Hao J, Yao Z, Tang Y, Remis A, Wu K, Yu X. Artificial Intelligence in Physical Therapy: Evaluating ChatGPT's Role in Clinical Decision Support for Musculoskeletal Care. *Ann Biomed Eng* 2025;53:(1)9-13. [\[CrossRef\]](#)
- Bilika P, Stefanouli V, Strimpakos N, Kapreli EV. Clinical reasoning using ChatGPT: Is it beyond credibility for physiotherapists use? *Physiother Theory Pract* 2024;40:(12)2943-62. [\[CrossRef\]](#)
- Deng J, Li L, Oosterhof JJ, Malliaras P, Silbernagel KG, Breda SJ, et al. ChatGPT is a comprehensive education tool for patients with patellar tendinopathy, but it currently lacks accuracy and readability. *Musculoskelet Sci Pract* 2025;76:103275. [\[CrossRef\]](#)
- Lobet S, Timmer M, Königs C, Stephensen D, McLaughlin P, Duport G, et al. The role of physiotherapy in the new treatment landscape for haemophilia. *J Clin Med* 2021;(13):2822. [\[CrossRef\]](#)
- Stephensen D, de Kleijn P, Matlary RED, Katzerova M, McLaughlin P, Ryan A, et al. Scope of practice of haemophilia physiotherapists: A European survey. *Haemophilia* 2019;25(3):514-50. [\[CrossRef\]](#)
- Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am* 2023;105(19):1519-26. [\[CrossRef\]](#)
- Johns WL, Kellish A, Farronato D, Ciccotti MG, Hammoud S. ChatGPT can offer satisfactory responses to common patient questions regarding elbow ulnar collateral ligament reconstruction. *Arthrosc Sports Med Rehabil* 2024;6(2):100893. [\[CrossRef\]](#)
- Artioli E, Veronesi F, Mazzotti A, Brogini S, Zielli SO, Giavaresi G, et al. Assessing ChatGPT responses to common patient questions regarding total ankle arthroplasty. *J Exp Orthop* 2024;12(1):e70138. [\[CrossRef\]](#)
- Pérez-Alenda S, Núñez-Cortés R, Calatayud J, Cruz-Montecinos C. Quality of life in hemophilia. In: *Handbook of the Behavior and Psychology of Disease*. Cham (CH): Springer Nature Switzerland; 2024. p. 1-15. [\[CrossRef\]](#)
- Srivastava A, Santagostino E, Dougall A, Kitchen S, Sutherland M, Pipe SW, et al. WFH guidelines for the management of hemophilia, 3<sup>rd</sup> edition. *Haemophilia* 2020;26(Suppl 6):1-158. [\[CrossRef\]](#)
- Forsyth AL, Zourikian N, Valentino LA, Rivard GE. The effect of cooling on coagulation and haemostasis: should "Ice" be part of treatment of acute haemarthrosis in haemophilia? *Haemophilia* 2012;18(6):843-50. [\[CrossRef\]](#)
- Guodemar-Pérez J, Ruiz-López M, Rodríguez-López E, García-Fernández P, Hervás-Pérez JP. Physiotherapy treatments in musculoskeletal pathologies associated with haemophilia. *Hamostaseologie* 2018;38(3):141-9. [\[CrossRef\]](#)
- de Kleijn P, Duport G, Jansone K, Marinić M, McLaughlin P, Noone D, et al. European principles of care for physiotherapy provision for persons with inherited bleeding disorders: Perspectives of physiotherapists and patients. *Haemophilia* 2022;28(4):649-55. [\[CrossRef\]](#)
- Truque-Díaz C, Meroño-Gallut J, Cuesta-Barriuso R, Pérez-Llanes R. Joint and myofascial manual therapy techniques in haemophilic ankle arthropathy: a randomized pilot study. *Haemophilia* 2025;31(2):295-303. [\[CrossRef\]](#)
- Tat NM, Can F, Sasmaz HI, Tat AM, Antmen AB. The effects of manual therapy on musculoskeletal system, functional level, joint health and kinesiophobia in young adults with severe haemophilia: A randomized pilot study. *Haemophilia* 2021;27(2):e230-e8. [\[CrossRef\]](#)
- Chimeno-Hernández A, Querol-Giner F, Pérez-Alenda S, Núñez-Cortés R, Cruz-Montecinos C, Carrasco JJ, et al. Effectiveness of physical exercise on postural balance in patients with haemophilia: A systematic review. *Haemophilia* 2022;28(3):409-21. [\[CrossRef\]](#)
- Cruz-Montecinos C, Núñez-Cortés R, Chimeno-Hernández A, López-Bueno R, Andersen LL, Mendez-Rebolledo G, Pérez-Alenda S, et al. Exercise variables and pain threshold reporting for strength training protocols in people with haemophilia: A systematic review of clinical trials. *Haemophilia* 2023;29(3):695-708. [\[CrossRef\]](#)
- Steen Carlsson K, Winding B, Astermark J, Baghaei F, Brodin E, Funding E, et al. Pain, depression and anxiety in people with haemophilia from three Nordic countries: Cross-sectional survey data from the MIND study. *Haemophilia* 2022;28(4):557-67. [\[CrossRef\]](#)
- Fenelon C, Murphy EP, Fahey EJ, Murphy RP, O'Connell NM, Queally JM. Total knee arthroplasty in hemophilia: survivorship and outcomes-a systematic review and meta-analysis. *J Arthroplasty* 2022;37(3):581-92.e1. [\[CrossRef\]](#)
- Deniz V, Guzel NA, Lobet S, Antmen AB, Sasmaz HI, Kilci A, et al. Effects of a supervised therapeutic exercise program on musculoskeletal health and gait in patients with haemophilia: A pilot study. *Haemophilia* 2022;28(1):166-75. [\[CrossRef\]](#)
- Hashem F, Stephensen D, Drechsler WI, Bladen M, Carroll L, Tracy PH, et al. Muscle strengthening intervention for boys with haemophilia: Developing and evaluating a best-practice exercise programme with boys, families and health-care professionals. *Health Expect* 2020;23(5):1350-11. [\[CrossRef\]](#)

22. Matlary RED, Grinda N, Sayers F, Versloot O, McLaughlin P; EAHAD Physiotherapists Committee. Promoting physical activity for people with haemophilia in the age of new treatments. *Haemophilia* 2022;28(6):885-90. [\[CrossRef\]](#)
23. Batorova A, Banchev A, Boban A, Brand B, Brinza M, Kotnik BF, Kiss C, Puras G, Rajnoch J, Zapotocka E. Physical Activity Awareness and Understanding of Treatment Protection Among People With Haemophilia and Their Caregivers in Central Europe. *Haemophilia* 2025;31(5):893-902. [\[CrossRef\]](#)
24. Kennedy M, Roche S, McGowan M, Singleton E, Elsheikh E, O'Donovan M, et al. Physical activity, physical fitness and cardiometabolic risk amongst adults with moderate and severe haemophilia. *Haemophilia* 2023;29(1):72-83. [\[CrossRef\]](#)
25. Cotino C, Pérez-Alenda S, Cruz-Montecinos C, López-Bueno R, Núñez-Cortés R, Suso-Martí L, et al. Barriers and facilitators of physical activity in adults with severe haemophilia: A qualitative study. *Haemophilia* 2023;29(5):1334-42. [\[CrossRef\]](#)
26. Stephensen D. Rehabilitation of patients with haemophilia after orthopaedic surgery: a case study. *Haemophilia* 2005;11(Suppl 1):26-9. [\[CrossRef\]](#)
27. Poyraz Isleyen T, Tarakci E, Leblebici G, Yeldan I, Zulfikar B. Comparison of efficiency of closed kinetic chain exercises versus proprioceptive exercises in improving balance and gait in people with hemophilia: protocol for a randomized controlled trial. *JMIR Res Protoc* 2025;14:e66770. [\[CrossRef\]](#)
28. McCarthy CJ, Berkowitz S, Ramalingam V, Ahmed M. Evaluation of an Artificial Intelligence Chatbot for Delivery of IR Patient Education Material: A Comparison with Societal Website Content. *J Vasc Interv Radiol* 2023;34(10):1760-8.e32. [\[CrossRef\]](#)
29. Van de Wyngaert C, Iarossi M, Hermans C. How good does ChatGPT answer frequently asked questions about haemophilia? *Haemophilia* 2023;29:1646-8. [\[CrossRef\]](#)
30. Zhu W, Geng W, Huang L, Qin X, Chen Z, Yan H. Who could and should give exercise prescription: Physicians, exercise and health scientists, fitness trainers, or ChatGPT? *J Sport Health Sci* 2024;13(3):368-72. [\[CrossRef\]](#)

**Appendix 1.** List of questions and ChatGPT's responses

Question 1: What are the treatment options other than pharmacological treatment for joint health in hemophilia?

ChatGPT Response:

1. Physiotherapy and Exercise: Strengthening muscles, flexibility exercises, balance and coordination training, and low-impact activities such as swimming, cycling, and walking.
2. Orthopedic Interventions: Bracing or splinting, orthotic devices, and surgical options (if joints are severely damaged) like synovectomy, joint replacement, or arthroscopic procedures.
3. Lifestyle Modifications: Weight management and avoiding high-risk activities (e.g., contact sports).
4. Education and Prevention: Teaching proper body mechanics and safe movement strategies. Early recognition of joint bleeds to treat them promptly.
5. Occupational Therapy: Helps patients adapt their daily activities to protect joints and maintain independence.
6. Complementary Therapies (used cautiously): Techniques like Tai Chi or Yoga.

Analysis: Satisfactory Response Requiring Mild Clarification

ChatGPT's response to non-pharmacological treatment options for joint health provides a comprehensive overview of the care in hemophilia. However, this answer does not provide information on which treatments to choose when considering the condition of people with hemophilia, nor does it include the advantages and disadvantages of these treatment options over each other. It did not offer any guidance on when orthopedic intervention would be preferred.

Question 2: What should I do for joint health in acute bleeding in hemophilia and can you create an emergency intervention guide in such a case?

ChatGPT Response: "It's critical to act fast and correctly to protect the joint long-term."

1. Immediate Actions ("R.I.C.E.") (See Table-1. Quick Checklist for Acute Joint Bleed in Hemophilia)
2. Factor Replacement Therapy: Administer clotting factor concentrate as soon as possible – ideally within 2 h. Dose depends on severity and type of hemophilia.
3. Early Phase Rehabilitation (first 24–72 h): Continue "R.I.C.E.". Pain management: Use prescribed medications if needed (avoid NSAIDs like ibuprofen unless advised – they can worsen bleeding). Monitor signs: Watch for increased swelling, redness, or inability to move.
4. Post-Acute Phase (after bleeding is controlled): Gentle Mobilization: Start pain-free range of motion exercises (under physiotherapist supervision). No weight-bearing until cleared. Gradual Strengthening: Progress to light strengthening exercises. Focus first on isometric exercises.
5. Long-Term Joint Protection: Restore full range of motion, strengthen muscles to prevent future bleeds, train for joint stability and proprioception, prevent Re-bleeding: Gradual return to daily activities and sport. Continue low-impact exercise.

Question 3: Which physiotherapy approaches are used to improve joint health in hemophilia?

ChatGPT Response: "Physiotherapy is a core part of joint care in hemophilia."

1. Range of motion (ROM) Exercises: Passive (therapist moves the joint) early if patient cannot move. Active (patient moves the joint) when pain/swelling allows. Example: Gentle bending and straightening of the elbow or knee.
2. Strengthening exercises: Start with isometric exercises. Progress to light resistance exercises once healing advances. Example: Quadriceps setting.
3. Proprioception and balance training: Standing on one leg. Balance boards or foam surfaces. Progress slowly to avoid micro-bleeds.
4. Functional training: Gait training. Step-ups, sit-to-stand exercises.
5. Manual therapy (when stable): Gentle joint mobilizations (only by trained therapists familiar with hemophilia). Note: Never during acute bleeding.
6. Hydrotherapy (aquatic therapy): Exercises done in warm water (buoyancy supports the body). Example: Water walking, gentle leg lifts.
7. Education and home exercise programs: Teach safe exercise techniques. Give customized home plans.

Key Principles for Physiotherapy

- Always treat bleeds first before starting exercises.
- Pain is a warning sign – exercise must be pain-free.
- Start slowly and gently, increase intensity only when safe.
- Work closely with a hematologist and physio specialized in hemophilia care.

## Appendix 1 (cont). List of questions and ChatGPT's responses

---

Question 4: Which manual therapy techniques are used to improve joint health in hemophilia?

ChatGPT Response: "Manual therapy can be used carefully in hemophilia, but only under certain conditions (no active bleeding, stable joint, trained therapist).

1. Soft tissue mobilization (massage): Gentle, slow strokes (effleurage). Deeper techniques (petrissage) only if there's no bleeding risk. Target areas: Muscles around affected joints (e.g., quadriceps, calf muscles).
2. Passive joint mobilizations: Grade I–II mobilizations (small amplitude, non-painful movements). Focus on accessory movements (glides, small oscillations). Example: Gentle knee joint gliding or ankle dorsiflexion mobilization.
3. Myofascial release: Very light, sustained pressure techniques over muscles and fascia. Important: Deep or aggressive release is avoided.
4. Manual stretching: Slow, passive stretches within a pain-free range. Example: Stretching hamstrings while keeping the knee slightly bent if needed.
5. Lymphatic drainage techniques: Very light, rhythmic strokes along lymphatic pathways. Used Especially After: Acute bleeds, to help resolve joint effusions.

Critical safety rules for manual therapy

- Never during acute bleeding or immediately after a large hemarthrosis
  - Always pain-free – pain signals possible micro-bleeding
  - Low-force, high-control movements
  - Close monitoring – stop if any swelling, warmth, or unusual pain appears
  - Must be done by a therapist experienced in hemophilia care.
- 

Question 5: What types of exercises are recommended to improve joint health in hemophilia?

Improving joint health in hemophilia is a multi-faceted approach, and exercise plays a crucial role in maintaining joint function and preventing long-term damage.

1. Range of motion (ROM) Exercises: Passive ROM (by therapist) in the early stages after a bleed if the joint is swollen or painful. Active ROM (self-movement) once swelling and pain subside. Example: Elbow flexion and extension (bending and straightening the arm). Ankle dorsiflexion and plantarflexion (moving the foot up and down).
2. Strengthening exercises: Isometric exercises (muscle contraction without joint movement) early in recovery. Gradually progress to concentric and eccentric strengthening (muscle shortening and lengthening) as the joint stabilizes. Example: Isometric quadriceps setting, leg raises, and resistance band exercises for upper and lower limbs to strengthen key muscles.
3. Proprioception and balance training: Work on balance by standing on one leg or using balance boards. Progress to more complex exercises like standing on foam surfaces or using a wobble board to challenge stability. Example: Standing on one leg for 20–30 s, alternating legs. Heel-to-toe walking.
4. Low-impact aerobic exercise: Use low-impact activities that avoid joint strain but provide cardiovascular benefits. Example: swimming, cycling, walking.
5. Stretching exercises: Gentle static stretching once the joint is pain-free (no aggressive stretching during a bleed or in acute recovery). Focus on muscle groups around the most affected joints. Example: Hamstring stretch (lying down, extend leg and gently pull toward the chest). Quadriceps stretch (standing, pulling foot toward the glutes).
6. Aquatic therapy (Hydrotherapy): Perform gentle strengthening, stretching, and aerobic exercises in water. Example: Walking or jogging in chest-deep water. Leg lifts or squats in water, where the buoyancy reduces the impact on the joints.
7. Functional movement training: Train movements required for daily tasks (walking, stair climbing, etc.) in a controlled manner. Gait training or using assistive devices if necessary. Example: Step-ups: To simulate walking up stairs. Sit-to-stand exercises: To mimic rising from a seated position.

Key principles for exercise

- Pain-Free: Exercises should never cause pain – pain could indicate micro-bleeding
  - Progress gradually: Start with low intensity and build up over time
  - Safety first: Avoid high-impact or risky activities (e.g., contact sports, running on hard surfaces)
  - Consistency: Regular exercise prevents muscle weakness and joint problems.
-

**Appendix 1 (cont).** List of questions and ChatGPT's responses

Question 6: If I do not receive physiotherapy treatment for joint health, what changes will occur in my body in the long term?

1. Joint damage (hemophilic arthropathy): Repeated joint bleeds can lead to chronic inflammation and cartilage damage over time. Without physiotherapy to strengthen muscles and maintain joint function, these changes can worsen. Effect: Joint deformities (e.g., flexion contractures, joint misalignment). Loss of joint space due to cartilage degradation. Severe pain and reduced range of motion.
2. Muscle weakness and atrophy: Without proper strengthening exercises, the muscles around affected joints weaken, leading to muscle atrophy (wasting). Weak muscles cannot support joints effectively, increasing the risk of further injury. Effect: Muscle weakness in areas like the thighs, calves, and upper arms. Increased risk of joint instability and frequent bleeds because muscles are not providing enough protection. Reduced overall mobility and difficulty performing daily activities.
3. Loss of range of motion (contractures): Without physiotherapy to maintain or restore flexibility, joints can become stiff, and surrounding tissues may shorten (leading to contractures). Effect: Loss of mobility in critical joints like elbows, knees, and ankles. Difficulty with basic movements such as bending, straightening, or walking. Permanent joint deformities if left unchecked.
4. Increased risk of future bleeding: As muscles weaken and joints become unstable, the risk of accidental falls or movements that lead to new joint bleeds increases. The lack of physiotherapy may prevent the proper muscle reinforcement that would otherwise protect joints. Effect: More frequent and severe bleeding episodes. A vicious cycle of repeated bleeds causing ongoing damage and dysfunction.
5. Chronic pain and inflammation: Without regular physiotherapy, the damaged joints and muscles remain unaddressed, causing chronic pain, swelling, and inflammation. Effect: Persistent joint pain that worsens with activity or weather changes. Inability to manage pain without consistent physical therapy to address underlying joint dysfunction.
6. Decreased quality of life: All of the above effects accumulate over time, reducing your ability to participate in daily activities, work, or recreational activities. The lack of mobility and persistent pain can lead to social isolation or depression. Effect: Inability to engage in exercise or low-impact activities (e.g., swimming, walking). Difficulty performing tasks such as household chores, climbing stairs, or even walking. Emotional and mental health struggles due to ongoing disability and pain.
7. Early onset of joint replacement needs: If joint health deteriorates significantly without physiotherapy, severe joint damage may lead to the need for joint replacement (e.g., knee or hip replacement) earlier than in those with regular joint care. Effect: Need for surgical interventions like joint replacements or synovectomies to manage pain and disability. A longer recovery process from these surgeries, especially if the body is weakened due to lack of physiotherapy.

Question 7: What are the potential risks of physiotherapy applications in hemophilic individuals?

Physiotherapy must be approached cautiously due to the unique risks involved. Improper or aggressive physiotherapy can lead to complications due to the bleeding tendency inherent in hemophilia.

1. Increased risk of bleeding: Hemophilic individuals are more prone to bleeding, even with minor trauma or intense physical activity. Aggressive or high-impact physiotherapy can cause soft tissue or joint bleeds. Example: Excessive stretching, mobilizations, or joint manipulations may cause small tears in tissues or increased joint pressure that can trigger a bleed.
2. Joint damage due to excessive force: High-velocity or aggressive manual therapy (e.g., deep joint mobilizations) or exercises involving high resistance can lead to joint trauma. Example: Improper techniques like high-impact stretching or sudden movements may worsen existing joint damage (hemophilic arthropathy).
3. Overloading the joint or muscle: Pushing too hard in strengthening exercises, especially if muscles are already weak from previous bleeds, can overload the joint and cause additional stress, potentially leading to micro-tears or joint damage. Example: Lifting too much weight or performing excessive repetitions during strengthening exercises can lead to strain or even more severe joint injuries.
4. Incorrect timing of exercise: Physiotherapy must be carefully timed in relation to bleeding episodes. Exercising too soon after a bleed (before swelling or pain is under control) can worsen the condition. Example: Early mobilization after a joint bleed, without allowing enough time for factor replacement and rest, could lead to further bleeding or joint instability.
5. Increased pain and swelling: Physiotherapy can sometimes irritate the tissues around a healing joint or muscle, especially if exercises are performed too intensely or too soon. Example: Too much stretching or joint mobilization can cause swelling, discomfort, or additional pain, especially in joints that are already inflamed or healing from a bleed.
6. Risk of overuse injuries: Repeating certain movements or exercises incorrectly (without proper supervision or technique) can lead to overuse injuries, particularly in weakened joints or muscles. Example: Repetitive strain from improper posture or poor biomechanics during exercises can increase the risk of injury or joint irritation, especially in those with pre-existing joint damage.
7. Lack of qualified supervision: Physiotherapy should ideally be performed under the supervision of a trained physiotherapist who is experienced in working with individuals with hemophilia. Inadequate supervision or incorrect modifications to exercises can lead to injuries or improper progression of treatment. Example: A physiotherapist unfamiliar with hemophilia might recommend exercises that are too aggressive or don't take the unique bleeding risks into account.

Preventive measures to minimize risks

1. Personalized exercise plans: Exercise programs must be tailored to the individual's bleeding history, joint condition, and overall health. Low-impact and joint-protective exercises are preferred (e.g., water therapy, cycling).
2. Slow and gradual progression: Any strengthening or stretching program should start slowly and progressively increase intensity, ensuring muscle recovery after each session.
3. Avoid high-impact or high-risk movements: Avoid activities that risk falls, bruises, or high joint stress (e.g., heavy lifting, high-intensity sports, sudden jerky movements).
4. Always monitor pain and swelling: Regular monitoring of pain levels and swelling during and after therapy is crucial. Any noticeable increase in these should be a signal to modify or stop the exercise.
5. Factor replacement before therapy: If a bleeding episode occurred, ensure proper factor replacement before beginning physiotherapy to reduce bleeding risks.
6. Frequent communication with medical providers: Keep open communication with the treating hematologist, as they can offer guidance on how to safely proceed with physiotherapy, including timing and intensity adjustments after bleeding events.

## Appendix 1 (cont). List of questions and ChatGPT's responses

Question 8: What should be the duration and frequency of physiotherapy for joint health in hemophilic individuals?

For individuals with hemophilia, maintaining joint health with physiotherapy is a lifelong need, not just a short treatment.

Duration:

- Continuous/lifelong: Physiotherapy should be part of their routine health care forever
- Specific sessions (like after a bleed, injury, or surgery) can last several weeks to months, depending on recovery
- Even when there is no active joint problem, ongoing maintenance exercises are important to prevent new bleeds and arthropathy.

Frequency:

- In stable periods (no acute bleed):
- Supervised physiotherapy: Typically 1–2 times/week (e.g., in a hospital or clinic) for assessments, correction of techniques, and adjustments.
- Home exercise program: Daily or 4–5 times/week, focusing on strength, flexibility, and proprioception.

After an acute joint bleed or surgery:

- Initially, more frequent: Physiotherapy can be daily or every other day early on, then taper as healing progresses.
- For children (especially): Include fun activities such as swimming, cycling, or controlled sports to improve adherence and promote joint health.

Important conditions:

- Always coordinate exercises with factor replacement therapy if needed (do physical activity when clotting factor levels are optimal)
- Avoid high-impact or risky activities that could cause joint trauma.

Summary: Lifelong daily exercises at home + 1–2 supervised physiotherapy sessions per week, adjusted during special periods like after bleeds or surgery.

---

Question 9: Does fitness or physical activity improve joint health in hemophilia?

Fitness and physical activity can significantly improve joint health in PwH, provided the activities are carefully chosen and adapted to avoid excess strain or injury. However, it is essential to choose low-impact activities and monitor intensity to avoid causing bleeding or joint damage.

1. Strengthening muscles around the Joints: Provides stability and reduces the stress placed on the joint itself, especially when the muscles are weak due to repeated bleeds or immobility.
2. Increasing joint mobility: Range of motion (ROM) exercises help maintain joint flexibility and prevent contractures and stiffness that can result from hemophilia-related joint damage.
3. Improving proprioception (joint awareness): Proprioception exercises (exercises that focus on balance and body awareness) improve the sense of joint position, which is crucial for preventing joint injuries and falls. This is especially important for preventing the risk of injury during everyday movements or physical activities.
4. Reducing the risk of further bleeding: Regular physical activity helps improve circulation and blood flow, which can support overall vascular health. This can reduce the likelihood of small bleeds in the joint (hemarthrosis) as the body becomes more efficient in managing circulation and repairing damage.
5. Pain reduction and improved function: Regular movement and exercise can help reduce joint pain by enhancing the fluid dynamics within the joint, improving muscle strength, and increasing mobility. Strengthening the muscles surrounding the joint also reduces the mechanical load on the joint itself, potentially decreasing pain over time.
6. Preventing obesity and overloading of joints: Weight management is essential for individuals with hemophilia. Excess weight puts additional strain on the joints, especially the lower body (hips, knees, and ankles), increasing the risk of joint wear and tear. Regular physical activity helps in maintaining a healthy body weight, thus reducing joint load.

Precautions

1. Avoid high-impact activities: Running, jumping, contact sports, or activities with a high risk of falls (e.g., basketball, soccer) should be avoided as they can lead to bleeding episodes.
  2. Start slowly: Begin with gentle movements, gradually increasing intensity and duration based on tolerance and joint condition.
  3. Pain-free movements: Ensure exercises are pain-free. Pain during or after exercise may indicate strain or a possible bleed.
  4. Monitor Swelling and Bruising: If any swelling, bruising, or discomfort occurs during or after an exercise session, adjust the intensity or take a break to allow recovery.
-

**Appendix 1 (cont).** List of questions and ChatGPT's responses

Question 10: When should physiotherapy be started after orthopedic surgery in hemophilia?

Physiotherapy is typically started as soon as it is safe to do so. The exact timing depends on the type of surgery, the severity of hemophilia, and the individual's recovery progress.

1. Immediately post-operative phase (0–48 h): Focus is on managing swelling, pain, and preventing complications such as joint stiffness or muscle atrophy. This period may involve gentle movements and techniques to ensure that the surgical site remains stable and protected.
2. Early rehabilitation (3–7 days): Physiotherapy can begin with light, passive range-of-motion exercises, and muscle activation if the surgeon agrees. Hemophilia patients need to be closely monitored to avoid bleeding or joint damage, so exercises must be carefully controlled.
3. After 2 weeks: If there are no significant complications, more active rehabilitation can begin, which may include strengthening exercises, weight-bearing activities, and improving flexibility. The intensity of the exercises should be progressively increased under careful supervision.

It is crucial to tailor the rehabilitation plan to the individual's condition, monitor for any bleeding episodes, and ensure that factor replacement therapy is managed to avoid complications. Always follow the advice of the treating orthopedic surgeon and hemophilia care team.

---

# Artificial Intelligence in Unexplained Infertility: A Systematic Review of Machine Learning-Based Predictive Models

 Tuğba Tahta,  Özge Aydoğan Aşir

Department of Midwifery, Ankara Medipol University Faculty of Health Sciences, Ankara, Türkiye

## Abstract

**Introduction:** This systematic review evaluates the methodological rigor and risk of bias of machine learning-based predictive models developed to estimate the success of assisted reproductive technologies in cases of unexplained infertility, using the Prediction Model Risk of Bias Assessment Tool.

**Methods:** A systematic review was conducted to assess predictive modeling studies focused on unexplained infertility and based on machine learning, guided by the framework of the Prediction Model Risk of Bias Assessment Tool. After rigorously screening 912 records, only three studies met the inclusion criteria. While limited in number, these studies highlight emerging evidence in this underexplored area.

**Results:** The included studies applied supervised machine learning algorithms, such as Random Forest, Support Vector Machines, Partial Least Squares Discriminant Analysis, and neural networks, across various biomedical data types. Reported predictive performance varied by data modality: spectroscopy-based models demonstrated high classification accuracy, ranging from 92% to 100%, while a couple-based metabolic model with external validation achieved an accuracy of 73.8%. According to the PROBAST assessment, two studies were rated as low risk of bias, whereas one study exhibited an unclear risk, primarily due to limitations in external validation and analytical transparency.

**Discussion and Conclusion:** This systematic review demonstrates the potential of machine learning-based models to enhance clinical decision-making in the context of unexplained infertility.

**Keywords:** Artificial intelligence; Infertility; Machine learning

According to the World Health Organization (WHO), infertility is defined as the inability to achieve pregnancy after one year (or more) of regular, unprotected sexual intercourse. Unexplained infertility is characterized by the inability to conceive despite regular, unprotected intercourse in the absence of any proven reproductive pathology in either partner.<sup>[1-3]</sup> Currently, there is no consensus on the diagnostic tests required to identify unexplained infertility.<sup>[3]</sup>

Various assisted reproductive technologies (ART) are commonly used for couples diagnosed with unexplained infertility, with *in vitro* fertilization (IVF) being the most preferred treatment approach.<sup>[4-6]</sup> However, ART is typically only partially subsidized and remains an expensive treatment protocol, imposing a substantial financial burden on couples. Reported pregnancy rates after ART generally range between 30% and 40%, highlighting the challenges of individualizing prognosis, the need for detailed clinical

**Cite this article as:** Tahta T, Aydoğan Aşir Ö. Artificial Intelligence in Unexplained Infertility: A Systematic Review of Machine Learning-Based Predictive Models. Lokman Hekim Health Sci 2026;6(2):290–302.

**Correspondence:** Tuğba Tahta, PhD. Ankara Medipol Üniversitesi, Sağlık Bilimleri Fakültesi, Ebelik Anabilim Dalı, Ankara, Türkiye

**E-mail:** tahtatugba@gmail.com **Submitted:** 09.12.2025 **Revised:** 20.12.2025 **Accepted:** 17.02.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



evaluations for each patient, and the cost-related barriers that add complexity to this issue.<sup>[7–9]</sup>

Artificial intelligence (AI), particularly its subfield of machine learning (ML), has demonstrated increasing efficacy in numerous studies aimed at improving clinical decision-support systems. ML algorithms are capable of identifying complex patterns and relationships within high-dimensional, heterogeneous biomedical data. In unexplained infertility—where the etiology is uncertain and data heterogeneity is high—ML offers a safe and robust alternative to traditional statistical approaches by providing enhanced predictive capability for IVF outcomes based on patient-specific characteristics. This allows clinicians to develop personalized treatment strategies and helps couples form more realistic expectations regarding pregnancy potential.<sup>[10–13]</sup>

A systematic review is needed to evaluate the quality, reliability, and risk of bias in existing ML models in this field. The Prediction Model Risk of Bias Assessment Tool (PROBAST) is a widely recognized framework for critically assessing predictive modeling studies.<sup>[14]</sup>

Despite the growing number of studies applying machine learning to reproductive medicine, a critical gap remains in the literature regarding unexplained infertility as a distinct clinical entity. Most ML-based infertility studies include heterogeneous populations, often combining multiple etiological subgroups without explicitly defining unexplained infertility or tailoring models to this diagnosis. As a result, the clinical validity and generalizability of these models for couples with unexplained infertility remain uncertain. Furthermore, existing reviews in this field primarily focus on algorithmic performance or technological advancements, while neglecting systematic evaluation of methodological rigor, risk of bias, and clinical applicability. To date, no systematic review has specifically assessed ML-based predictive models for unexplained infertility using a validated risk-of-bias framework. Addressing this gap is essential to determine whether current ML models provide reliable and clinically meaningful support for decision-making in this challenging and costly patient population.

The aim of this systematic review is to evaluate studies employing ML-based predictive models designed to forecast ART success in couples with unexplained infertility, using the PROBAST tool to assess methodological quality and provide insights for future research.

## Materials and Methods

This review was developed in accordance with a systematic review protocol, following the PRISMA-P 2015 checklist.<sup>[15]</sup>

The reliability and quality of the predictive models were assessed using PROBAST criteria,<sup>[14–16]</sup> ensuring transparent and systematic reporting of findings.

This systematic review was prospectively registered with PROSPERO (registration number: CRD420251037469). The reporting of this review was also conducted in line with the PRISMA-P 2015 statement for systematic reviews, particularly regarding the transparent reporting of the search strategy and study selection process.

## Research Question

In this study, the research question was structured based on the Population, Intervention, Comparison, Outcome, Study Design (PICOS) framework.<sup>[17]</sup>

The research question in this systematic review and/or meta-analysis was structured according to the PICOS framework:

- P (Population): The target population includes women, or heterosexual couples diagnosed with unexplained infertility.
- I (Intervention): The intervention involves the use of ML algorithms for predictive or diagnostic purposes related to infertility treatment outcomes.
- C (Comparison): The performance of ML models is compared to conventional diagnostic or predictive methods, or to other ML models when applicable.
- O (Outcomes): Primary outcomes include the prediction accuracy of pregnancy outcomes, accuracy of embryo selection, diagnostic accuracy, AUC, sensitivity, specificity, and other relevant performance metrics.
- S (Study Design): The included studies consist of randomized controlled trials, non-randomized quasi-experimental studies, correlational/observational studies, and descriptive studies, all of which were reported in a consistent manner.

The specific questions examined in this review are as follows:

- Research Question 1: What are the differences in terms of accuracy, sensitivity, and specificity between machine learning-based predictive models and traditional statistical methods in predicting pregnancy rates among individuals diagnosed with unexplained infertility?
- Research Question 2: What are the AUC values of different machine learning algorithms (e.g., Random Forest, SVM, Artificial Neural Networks, Deep Learning) in terms of embryo selection accuracy?

- Research Question 3: Do machine learning-based predictive models developed for cases of unexplained infertility possess sufficient reliability to be used confidently in clinical decision-making processes?

### Inclusion Criteria

- Original research studies utilizing ML or AI algorithms.
- Studies specifically addressing unexplained infertility, IVF, embryo selection, and/or treatment success.
- Research conducted using clinical data (i.e., real patient and/or embryo data).
- Studies published in peer-reviewed journals from the year 2015 onwards.

### Exclusion Criteria

- Studies presenting only theoretical models (i.e., simulation-based studies without clinical data).
- Studies relying solely on traditional statistical methods without the use of ML/AI techniques.
- Animal experiments or *in vitro* cell culture studies.
- Case reports, review articles, editorials, or commentary papers.

### Literature Search / Search Strategy

A comprehensive literature search was conducted in the electronic databases PubMed, Web of Science, Scopus, Embase, Google Scholar, and the Cochrane Library. All databases were searched from January 1, 2010, to January 1, 2025, which was the date of the final search update. The search included articles published in Turkish and English within the last 15 years encompassing randomized controlled trials, non-randomized quasi-experimental studies, correlational/observational studies, and descriptive studies. Titles and abstracts of the identified articles were screened according to the inclusion criteria. Relevant studies were imported into an EndNote library for further selection and categorization. MeSH terms and free-text keywords were used to formulate the search strategy. The following terms and their combinations were employed: "unexplained infertility," "infertility," "machine learning," "predictive modeling," "IVF," "pregnancy outcomes," "embryo selection," "artificial intelligence in infertility," "fertility prediction," "supervised learning," "reproductive outcomes," "randomized controlled trials," "non-randomized quasi-experimental studies," "correlational/observational studies," and "descriptive studies."

For each database, a database-specific full Boolean search string combining controlled vocabulary (e.g., MeSH terms)

**Table 1.** Summary of core search concepts (keywords used in the electronic search)

Keywords	
	Unexplained infertility
	Infertility
	Machine learning
	Predictive modeling
	IVF
	Pregnancy outcomes
	Embryo selection
	Randomized
	Randomized Controlled Trial
	Randomized Clinical Trial

IVF: *In vitro* fertilisation.

and free-text keywords was developed and adapted to the syntax of that database (e.g., field tags, truncation symbols, and operators). A summary of the core search concepts is presented in Table 1, whereas the complete electronic search strategies (including full Boolean strings for each database, with all field tags and limits) are provided in Appendix 1, in accordance with PRISMA-P 2015 recommendations for transparent reporting of search strategies.

### Study Selection

All records identified through database searches (n=912) were imported into EndNote for reference management and deduplication. After removing duplicates (n=101), 811 records were screened by title and abstract, and 785 were excluded, primarily due to keyword/indexing incompatibility with the review focus (n=570) or publication in other languages (n=215). The full texts of 26 reports were assessed for eligibility and all were successfully retrieved. Of these, 23 were excluded at the full-text stage for predefined reasons, most commonly because the study population was not explicitly restricted to unexplained infertility or because the unexplained infertility subgroup could not be isolated from mixed-etiology cohorts. Additional reasons for exclusion included different clinical conditions (e.g., PCOS/PCOD or recurrent pregnancy loss), outcomes not aligned with the review objectives, non-original article types, or absence of ML/AI-based predictive modeling. Ultimately, three studies met the inclusion criteria and were included in the synthesis (Fig. 1; Appendix 2).

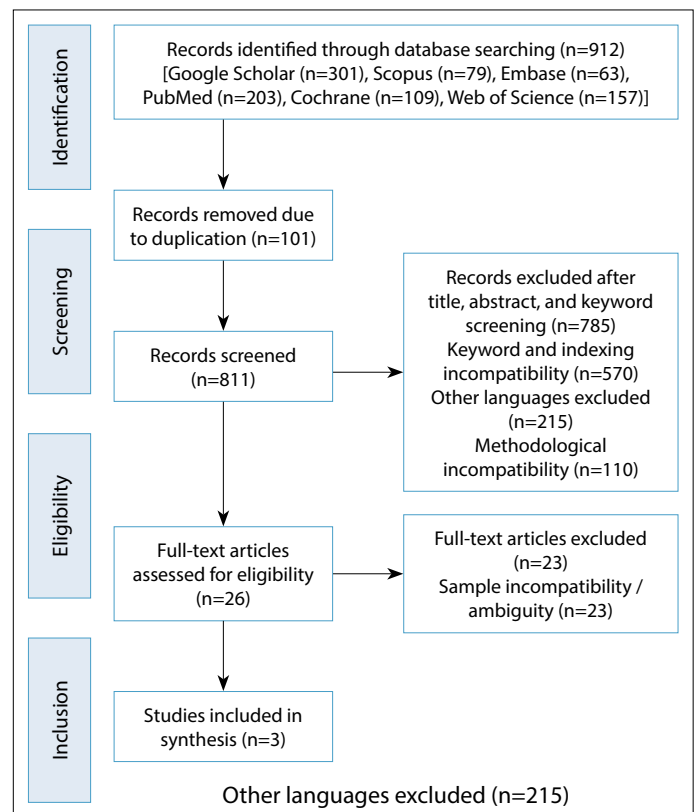
Despite an extensive and systematic search across multiple databases, only three studies met the strict inclusion criteria of this review. This limited yield is primarily attributable to the narrow clinical definition of unexplained infertility, which is often inconsistently reported or entirely omitted in the

broader machine learning–based infertility literature. Most ML studies in reproductive medicine analyze infertility as a heterogeneous condition or combine multiple etiological subgroups without explicitly identifying unexplained infertility as a distinct diagnostic entity, rendering them unsuitable for disease-specific synthesis. In addition, many ML-based models focus on embryo selection, semen analysis, oocyte quality, or general IVF outcomes such as clinical pregnancy, rather than on the diagnosis or classification of unexplained infertility, which was the targeted focus of this review. Therefore, only studies that explicitly identified unexplained infertility as a primary diagnostic criterion and reported predictive modeling outcomes for this specific population were included, while studies that investigated infertility without distinguishing between diagnostic subgroups were excluded to maintain clinical and methodological coherence. Furthermore, the application of strict methodological criteria—limiting inclusion to peer-reviewed primary research involving human participants and explicit ML-based predictive modeling—further reduced the eligible pool in accordance with the PICOS framework. Consequently, only three primary research articles fulfilled all inclusion criteria and were eligible for full-text synthesis. This limited number of included studies reflects a genuine gap and the early stage of ML applications in the specific field of unexplained infertility rather than limitations of the search strategy, as illustrated in the PRISMA flow diagram.

In addition to peer-reviewed journal articles, gray literature sources—including conference proceedings, dissertations, preprints, and non-indexed reports—were also screened to minimize publication bias. Searches were performed in Google Scholar, ProQuest Dissertations & Theses, and institutional repositories using the same keyword combinations. However, none of the gray literature sources met the predefined inclusion criteria (i.e., machine learning-based predictive modeling specifically targeting unexplained infertility in human participants). Therefore, although gray literature was thoroughly reviewed, no additional eligible studies were identified.

In addition to this, a detailed list of studies included and excluded at the full-text screening stage, together with brief reasons for exclusion, is presented in Appendix 2 to further enhance the transparency and reproducibility of the study selection process.

Most of the excluded articles addressed infertility in a general context, did not specifically distinguish unexplained infertility as a diagnostic category, or did not utilize machine learning-based predictive modeling methods. The



**Figure 1.** PRISMA flow diagram.

screening process was conducted independently by two reviewers (T.T., Ö.A.A.). Each reviewer assessed the titles, abstracts, and full texts according to predefined inclusion and exclusion criteria. Discrepancies between reviewers were resolved through discussion, and when necessary, a third reviewer was consulted to reach a consensus.

### Primary Outcomes

- Predictive accuracy of ML/AI models for clinical outcomes in unexplained infertility, particularly:
  - Pregnancy prediction accuracy
  - Embryo selection accuracy
  - Diagnostic accuracy

Performance metrics reported in the included studies, such as:

- AUC
- Sensitivity (true positive rate)
- Specificity (true negative rate)
- Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

Timing of outcome measurement: at the conclusion of the IVF treatment cycle or as defined by each study (e.g., clinical pregnancy, ongoing pregnancy, live birth).

**Table 2.** PROBAST risk of bias assessment – Depciuch et al.<sup>[18]</sup>

Domain	Signaling questions	Assessment	Comments
Participants	Were appropriate data sources used?	Low risk	Participants were selected from an IVF center with a diagnosis of unexplained infertility.
	Were inclusion/exclusion criteria appropriate?	Low risk	Clear criteria focusing on unexplained infertility cases.
Predictors	Were predictors defined and assessed appropriately?	Low risk	Raman spectral features were systematically measured and analyzed.
	Were predictor assessments blinded to outcome data?	Unclear risk	The study does not specify blinding procedures.
Outcome	Was the outcome defined and determined appropriately?	Low risk	Outcomes were based on established oxidative stress markers and spectral analysis.
	Was the outcome determined without knowledge of predictor information?	Unclear risk	Blinding of outcome assessment is not detailed.
Analysis	Were appropriate statistical methods used?	Low risk	Utilized multiple machine learning algorithms with high classification accuracy.
	Was model overfitting avoided?	Low risk	High accuracy suggests robustness, but external validation is not mentioned.
	Were performance measures appropriate?	Low risk	Accuracy metrics were reported for model performance.

PROBAST: Prediction Model Risk of Bias Assessment Tool.

Effect measure: Standardized reporting of model performance (e.g., AUC, accuracy, sensitivity, specificity), extracted as reported in each study for synthesis.

### Risk of Bias

The methodological quality and risk of bias of the included studies were evaluated using the PROBAST.<sup>[14]</sup> This tool is designed to assess the risk of bias and concerns regarding the applicability of studies that develop, validate, or update prediction models.

PROBAST consists of four key domains:

To provide greater methodological transparency and reproducibility, the PROBAST evaluation criteria were applied and reported in detail. PROBAST comprises 20 signaling questions organized into four core domains, each addressing potential sources of bias in prediction model studies:

- **Participants (P):** This domain assesses whether the study population is representative of the target population for the intended clinical use of the prediction model. Specific considerations include the data source (e.g., single-center vs. multicenter), recruitment strategy, eligibility criteria, and whether participant selection could introduce selection bias.
- **Predictors (Pr):** The predictors domain evaluates whether all candidate predictors were clearly defined, measured in a consistent and reproducible manner across participants, and assessed without knowledge of outcome status. Particular attention was paid to biological plausibility, measurement timing, and

potential information leakage between predictors and outcomes.

- **Outcome (O):** This domain examines whether the outcome was explicitly defined using clinically accepted criteria, measured objectively, and determined independently of predictor information. The timing of outcome assessment and the presence or absence of blinding were also considered, as these factors may influence outcome misclassification.
- **Analysis (A):** The analysis domain focuses on the statistical rigor of model development and validation. This includes evaluation of sample size adequacy, handling of missing data, prevention of model overfitting, use of appropriate internal and/or external validation strategies, transparency of model specification, and the appropriateness of reported performance metrics (e.g., accuracy, sensitivity, specificity, AUC).

Each signaling question was independently rated by two reviewers as “Yes,” “Probably Yes,” “Probably No,” “No,” or “No Information,” in accordance with PROBAST guidance. Domain-level risk of bias judgments (Low Risk, High Risk, or Unclear Risk) were derived by aggregating responses within each domain, with particular emphasis on the analysis domain due to its critical role in prediction model validity. In addition to risk of bias, applicability concerns were evaluated separately for the Participants, Predictors, and Outcomes domains to determine the clinical relevance of each study to the review question focusing on unexplained infertility. Any disagreements between reviewers were

**Table 3.** PROBAST Risk of Bias Assessment – Jakubczyk et al.<sup>[19]</sup>

Domain	Signaling questions	Assessment	Comments
Participants	Were appropriate data sources used?	Low risk	Participants were recruited from a fertility clinic and included idiopathic infertility cases and fertile controls.
	Were inclusion/exclusion criteria appropriate?	Low risk	Inclusion and exclusion criteria were clearly described and appropriate for the research question.
Predictors	Were predictors defined and assessed appropriately?	Low risk	FTIR spectral features and gonadotrophin levels were predefined and biologically plausible predictors.
	Were predictor assessments blinded to outcome data?	Unclear risk	It is not explicitly stated whether spectral data analysis was blinded to infertility status, which may introduce bias.
Outcome	Was the outcome defined and determined appropriately?	Low risk	Idiopathic infertility diagnosis was made using standard clinical criteria.
	Was the outcome determined without knowledge of predictor information?	Low risk	Outcome classification was based on clinical records and unlikely to be influenced by predictor data.
Analysis	Were appropriate statistical methods used?	Some concerns	While multivariate and machine learning methods (PLS, SVM) were used, details on validation strategy were limited.
	Was model overfitting avoided?	Some concerns	External validation was not clearly reported; risk of overfitting exists due to potential model complexity.
	Were performance measures appropriate?	Low risk	Model performance was presented (sensitivity, specificity), though external validation was limited.

PROBAST: Prediction Model Risk of Bias Assessment Tool.

**Table 4.** PROBAST Detailed Risk of Bias Table – Bachelot et al.<sup>[20]</sup>

Domain	Signaling questions	Assessment	Comments
Participants	Were appropriate data sources used?	Low risk	Participants were recruited from a clinical fertility cohort; both infertile and fertile couples clearly defined.
	Were inclusion/exclusion criteria appropriate?	Low risk	Inclusion/exclusion criteria were explicitly reported; infertility defined as >12 months TTP.
Predictors	Were predictors defined and assessed appropriately?	Low risk	Anthropometric, metabolic, and antioxidative variables were well-described and clinically relevant.
	Were predictor assessments blinded to outcome data?	Low risk	Predictor measurement was independent of fertility status; outcome knowledge unlikely to bias assessment.
Outcome	Was the outcome defined and determined appropriately?	Low risk	Outcome (fertile vs. idiopathic infertility) was based on standard clinical definitions and timeline to pregnancy.
	Was the outcome determined without knowledge of predictor information?	Low risk	Outcomes were historical or prospectively collected; unlikely to be influenced by predictor values.
Analysis	Were appropriate statistical methods used?	Low risk	Random Forest and OPLS-DA were appropriate for classification; external validation was performed.
	Was model overfitting avoided?	Low risk	The model was trained on a development set and tested on a separate external validation set.
	Were performance measures appropriate?	Low risk	Accuracy was reported; variables were reduced to 13 features, and performance validated externally.

PROBAST: Prediction Model Risk of Bias Assessment Tool.

resolved through discussion until consensus was reached. The detailed results of the PROBAST assessment are presented in Tables 2, 3, 4, with an overall summary provided in Table 5 and a narrative synthesis included in the Results section.<sup>[14–16]</sup>

## Synthesis Method

In this systematic review, statistical meta-analysis was not conducted due to the limited number of included primary studies (n=3). The primary reason for this decision was the lack of sufficient homogeneous data across studies to support

**Table 5.** Study characteristics and summary of PROBAST risk of bias assessment

Study	Participants risk	Predictors risk	Outcome risk	Analysis risk	Overall PROBAST risk
Depciuch et al. <sup>[18]</sup>	Low risk	Low/unclear risk	Low/unclear risk	Low risk	Unclear risk
Jakubczyk et al. <sup>[19]</sup>	Low risk	Low/unclear risk	Low risk	Unclear risk	Unclear risk
Bachelot et al. <sup>[20]</sup>	Low risk	Low risk	Low risk	Low risk	Low risk

A PROBAST scoring matrix was added to provide a visual summary of the risk-of-bias judgments across all four domains for each included study. The matrix illustrates domain-specific ratings (low, high, unclear) in a structured format, allowing readers to easily compare methodological strengths and limitations across studies. This matrix complements the narrative assessment and ensures transparency and reproducibility of the risk-of-bias evaluation.

meta-analysis. Instead, a narrative synthesis approach was adopted. The data extracted from each study were systematically organized in a summary table, considering factors such as the study's objectives, sample size and characteristics, machine learning algorithms used, targeted outcome variables (e.g., pregnancy prediction accuracy, embryo selection accuracy, diagnostic accuracy), and performance metrics (AUC, sensitivity, specificity, accuracy rate, etc.). Similarities and differences among the studies were comparatively analyzed under thematic headings.

Furthermore, to assess the methodological quality and potential bias risk of the predictive models developed in the studies, the PROBAST tool was utilized. PROBAST provides a systematic evaluation of the risk of bias across four domains (participants, predictors, outcomes, and analysis), offering insight into the reliability and generalizability of the models. This approach has made a significant contribution to deeply investigating the current state of machine learning applications for unexplained infertility and identifying gaps in the existing literature.<sup>[14]</sup>

## Results

A total of 3 studies meeting the inclusion and exclusion criteria were included in this systematic review. The methodological quality and risk of bias of the studies were assessed using the PROBAST criteria for risk of bias and applicability. Individual quality and risk assessments for each study, as well as the overall risk of bias summary table, are presented below.

### PROBAST Evaluation of Included Studies

This study demonstrated a low risk of bias across all PROBAST domains.<sup>[18]</sup> Participants were clearly defined, and the predictors—Raman spectral features and oxidative stress markers—were measured with robust methodology. A variety of machine learning algorithms with internal validation were applied, and outcome assessments were consistently conducted. Appropriate performance metrics were reported, supporting the model's predictive reliability. Overall, this study provides strong methodological rigor

and offers reliable evidence for the diagnostic application of machine learning in the context of unexplained infertility.

Overall PROBAST Risk of Bias Judgment: Unclear Risk.

This study included well-characterized participants and clinically relevant predictors, such as FTIR spectra and gonadotrophin levels.<sup>[19]</sup> However, there are concerns regarding the analysis domain. Specifically, the absence of explicit external validation and the potential lack of blinding during predictor assessment introduce possible biases. Therefore, while the findings are promising, the risk of bias is considered moderate or unclear. Further external validation is recommended to enhance the generalizability and practical utility of the model.

Overall PROBAST Risk of Bias Judgment: Unclear Risk.

This study demonstrated a low risk of bias across all PROBAST domains.<sup>[20]</sup> The predictors were biologically meaningful, clinically relevant, and measured appropriately. Outcome definitions were clearly established and assessed independently of predictor data. A couple-based modeling approach was used, and external validation was performed with consistent performance metrics. These factors collectively support the methodological robustness and clinical relevance of the model for predicting unexplained infertility.

Overall PROBAST Risk of Bias Judgment: Low Risk.

The three studies included in this systematic review were critically analyzed using a classical systematic review table format. This approach allowed us to systematically compare the methodology, participants, and key findings of each study. The studies were selected based on their use of ML or AI algorithms to predict outcomes in unexplained infertility, specifically focusing on clinical data related to embryo selection, pregnancy outcomes, or diagnostic accuracy. Below is a summary of the studies in the systematic review table.

### Analysis of the Included Studies

This systematic review examined three original studies that explored the application of ML techniques in the diagnosis and classification of unexplained female infertility. These

**Table 6.** Systematic review table (classic format)

Study	Design	Population	Sample size	Methods	Outcomes	Key findings
Depciuch et al., <sup>[18]</sup> Türkiye	Retrospective observational study	Women diagnosed with unexplained infertility undergoing IVF	128 follicular fluid samples (65 unexplained infertility, 63 controls)	Raman spectroscopy of follicular fluid; oxidative load measurements; multivariate analysis; six machine learning algorithms (Random Forest, C5.0 Decision Tree, k-NN, Deep Neural Networks, SVM, XGBoost)	Differences in Raman spectral bands associated with oxidative load, amide III, and lipids; classification accuracy of machine learning models	Significant spectral differences between groups; machine learning models achieved classification accuracy ranging from 92.04% to 99.17%
Jakubczyk et al., <sup>[19]</sup> Türkiye	Prospective observational study	Women diagnosed with idiopathic female infertility (IFI) undergoing IVF treatment at Academic Hospital IVF Center	116 women: 58 with idiopathic female infertility and 58 fertile controls; follicular fluid samples collected during IVF procedures	Fourier-transform infrared (FTIR) spectroscopy combined with multivariate analysis and six machine learning algorithms (e.g., Random Forest, SVM)	Identification of FTIR spectral characteristics associated with ovarian reserve and reproductive hormone levels; differentiation between IFI and control groups using multivariate analysis and machine learning methods; classification accuracy ranging from 93.75% to 100% depending on the algorithm used.	FTIR spectra revealed significant differences in lipid and protein vibrations between groups; machine learning models achieved classification accuracies ranging from 93.75% to 100% depending on the algorithm used
Bachelot et al., <sup>[20]</sup> France	Multicenter cross-sectional case-control study	Couples with primary idiopathic infertility (>12 months) and fertile couples with spontaneous conception (<12 months to pregnancy)	197 couples (97 infertile, 100 fertile); development set: 136 couples; test set: 61 couples	Machine learning model (Random Forest) based on couple-level data. Model was trained on 136 couples and tested on 61 couples	Development of a couple-based machine learning model (OPLS-DA) to stratify infertile and fertile couples based on anthropometric, metabolic, and antioxidative parameters; refined model with 13 variables achieved 73.8% accuracy in external validation; couple-based approach outperformed models considering men and women separately.	The model identified 13 predictive variables and achieved 73.8% accuracy in classifying idiopathic infertility. Couple-based prediction outperformed individual-based models.

studies utilized biochemical and spectroscopic analyses—primarily of follicular fluid—and applied a range of ML algorithms to identify potential biomarkers and enhance diagnostic precision.

In a retrospective, cross-sectional, observational study conducted,<sup>[18]</sup> a total of 128 follicular fluid samples were analyzed, including 65 from women with unexplained infertility and 63 from fertile controls. The samples were

evaluated using Raman spectroscopy, and the spectral data were analyzed via multivariate statistical approaches and correlation tests to assess oxidative stress markers. Six machine learning algorithms were employed for classification purposes: Random Forest, C5.0 decision tree, k-nearest neighbors (k-NN), deep neural networks, support vector machines (SVM), and XGBoost. The results indicated significant differences in oxidative stress-related

Raman bands, particularly within the amide III and lipid regions, between the unexplained infertility group and the control group. Classification accuracy for the machine learning models ranged between 92.04% and 99.17%. This study demonstrates that Raman spectroscopy combined with machine learning approaches can be utilized to identify potential biomarkers in the diagnosis of unexplained infertility. The findings suggest that oxidative stress-associated molecular changes in follicular fluid may contribute to a better understanding of the etiology of infertility and aid in diagnostic strategies.

In this cross-sectional, analytical observational study conducted,<sup>[19]</sup> a total of 116 follicular fluid samples (58 from women with unexplained female infertility (IFI) and 58 from fertile controls) were analyzed. The study utilized Fourier Transform Infrared (FTIR) spectroscopy to assess biochemical differences in follicular fluid, in combination with gonadotrophin levels. Spectral data were processed using multivariate statistical techniques including Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis (PLS-DA), along with machine learning algorithms such as Support Vector Machines (SVM) and Random Forest to classify the samples. The results revealed increased lipid absorption and decreased amide I and II bands in the IFI group, suggesting altered biochemical composition in the follicular environment. These alterations were associated with impaired reproductive potential. The applied machine learning models achieved classification accuracies up to 100%, demonstrating their diagnostic potential. This study highlights the promising role of FTIR spectroscopy combined with machine learning in identifying potential biomarkers for the diagnosis of unexplained infertility. The molecular alterations observed in follicular fluid may provide insight into biochemical mechanisms affecting oocyte quality and fertilization capacity.

In this prospective observational study conducted,<sup>[20]</sup> couples with primary unexplained infertility (failure to conceive after >12 months) were compared to fertile couples who achieved spontaneous conception within 12 months. A total of 197 couples were included (97 infertile and 100 fertile); 136 couples were used in the development dataset and 61 couples in the external test set. The researchers developed a machine learning model using Random Forest and Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) based on couple-level anthropometric, metabolic, and antioxidative parameters. A refined model incorporating 13 predictive variables achieved an external validation accuracy of

73.8%. Importantly, the couple-based predictive model significantly outperformed models that assessed men and women separately. This study demonstrates that integrating couple-level metabolic and biochemical data into machine learning models can effectively classify unexplained infertility. The findings support the clinical relevance of adopting a couple-based approach, rather than individual-based assessments, for improved diagnostic and predictive accuracy in infertility management.

### Types of Features and Predictors Used

The reviewed studies demonstrated variation according to data sources.

In spectroscopy-based studies,<sup>[18,19]</sup> predictor variables primarily included oxidative stress markers, Raman and FTIR spectral bands derived from follicular fluid, and gonadotropin levels—serving as biochemical and molecular indicators of oocyte quality and the follicular environment. In contrast, the study conducted<sup>[20]</sup> analyzed couple-level anthropometric data (e.g., BMI), metabolic markers (e.g., glucose and lipid levels), and antioxidant parameters. This systemic approach provided a broader and more holistic view of infertility-related physiology.

### Discussion

This systematic review examined three original studies applying ML techniques to the diagnosis and classification of unexplained female infertility. Among the three predefined research questions, the first and second were addressed through comparative evaluation of accuracy and AUC across studies, while the third—clinical reliability—was discussed in the context of the PROBAST assessment. However, due to the limited number of studies, some aspects could only be addressed narratively rather than quantitatively. The findings consistently indicate that integrating biochemical, metabolic, and spectroscopic data with ML approaches can significantly enhance diagnostic precision, potentially uncovering hidden pathophysiological mechanisms behind unexplained infertility. The selected studies demonstrate a strong topical alignment, all focusing on the application of machine learning methods to detect biomarkers or physiological indicators associated with unexplained infertility. They address critical gaps in understanding the etiology of unexplained infertility by employing advanced computational tools.<sup>[18–20]</sup>

The reviewed studies were published between 2021 and 2023. All three utilized observational study designs: two were cross-sectional analytical studies,<sup>[18,19]</sup> and one was

a prospective observational study.<sup>[20]</sup> The studies utilized a range of data types. Spectroscopy-based studies<sup>[18,19]</sup> focused on biochemical features extracted from follicular fluid samples, including Raman and FTIR spectral bands, oxidative stress markers, and gonadotrophin levels. In contrast,<sup>[20]</sup> employed anthropometric, metabolic, and antioxidative biomarkers at the couple level, providing a systemic physiological view. The primary objectives across studies were to enhance the diagnostic capabilities for unexplained infertility by applying ML methods to biomedical datasets, identify potential biomarkers, and improve understanding of the biological underpinnings of unexplained infertility. Notably,<sup>[20]</sup> extended the scope to consider couple-based features, emphasizing the dyadic nature of infertility. Although traditional supervised ML algorithms dominated, the inclusion of deep neural networks<sup>[18]</sup> represents a notable foray into deep learning approaches. However, there remains a limited application of unsupervised learning techniques across the reviewed studies. Future research should consider greater utilization of unsupervised and deep learning methods to discover latent structures in the data without predefined labels. While current studies predominantly focused on biochemical and metabolic predictors, the unexplored psychological or hormonal dimensions of unexplained infertility represent an important frontier.

Unexplained infertility imposes a profound emotional toll that extends beyond medical diagnosis. Due to the absence of a clear explanation for fertility difficulties, couples often experience anxiety, psychological strain, and frustration, which negatively affect their quality of life.<sup>[2,3,21,22]</sup> In addition to psychological distress, infertility treatments are associated with a substantial financial burden, further compounding emotional strain.<sup>[21–25]</sup> In this context, minimizing diagnostic uncertainty and error through advanced AI-based decision-support systems is of critical importance, as inaccurate or delayed diagnoses may amplify both psychological and financial consequences for affected couples.<sup>[26,27]</sup> However, existing ML-based studies in unexplained infertility remain limited by methodological constraints, including small sample sizes and restricted generalizability.<sup>[14]</sup>

Although reducing the psychological and financial burden of unexplained infertility represents an important long-term objective of ML-based decision-support systems, current evidence suggests that existing models are still far from achieving this goal in a clinically meaningful manner. The reviewed studies primarily rely on biochemical, spectroscopic, or metabolic markers,

while none incorporate patient-reported psychological distress, psychosocial variables, or cost-related outcomes. Moreover, the high diagnostic accuracy reported in spectroscopy-based studies (92–100%) is derived from small, single-center datasets without external validation, limiting their immediate applicability to routine clinical practice. Thus, while these early models may help shorten the diagnostic process and reduce uncertainty, their current level of development does not yet support direct translation into real-world clinical workflows. Future ML models should therefore be externally validated, integrated with broader biopsychosocial data, and evaluated prospectively to determine whether they can meaningfully improve patient experience and inform more cost-effective care pathways.

The comparative synthesis of the included studies demonstrates that machine learning models have strong potential in detecting biochemical and metabolic signatures associated with unexplained infertility; however, their current applicability is constrained by methodological and data-related factors. Among the algorithms used, spectroscopy-based datasets (Raman and FTIR) showed the highest predictive performance, with classification accuracies ranging from 92% to 100%, particularly when analyzed using Random Forest, SVM, and PLS-DA. These strong results are likely attributable to the high-dimensional molecular features contained in spectroscopic data, which are well suited to ML-based classification. In contrast, the couple-level metabolic dataset demonstrated more modest performance, achieving an externally validated accuracy of 73.8%. This suggests that systemic physiological markers provide clinically relevant but less discriminative signals than direct biochemical profiling of follicular fluid.

When examined through PROBAST, spectroscopy-based studies demonstrated a comparatively lower risk of bias in the predictor domain, as predictors were derived from standardized laboratory measurements. However, these studies showed higher risk of bias in the “participants” and “analysis” domains, primarily due to small sample sizes and the absence of external validation. Conversely, the couple-based study exhibited a lower risk of bias in the participant domain, supported by a larger dataset and the use of an independent test set. Nonetheless, it showed a higher risk of bias in the predictor domain, as metabolic and anthropometric variables may be influenced by unmeasured lifestyle or environmental confounders.

Taken together, three major barriers currently limit the translation of ML-based models for unexplained infertility into routine clinical practice:

- (1) Small, single-center datasets that may inflate performance estimates and limit generalizability;
- (2) Heterogeneity and lack of standardized outcome definitions, particularly regarding biochemical, clinical, and ongoing pregnancy; and
- (3) Limited external and prospective validation, which is essential before clinical implementation.

A deeper methodological comparison reveals important differences in model development and validation strategies. Spectroscopy-based studies<sup>[18,19]</sup> relied on high-dimensional spectral data but did not employ robust external validation, increasing the risk of overfitting despite their high reported accuracies. In contrast, Bachelot et al.<sup>[20]</sup> incorporated clinically interpretable predictors and applied true external validation, resulting in lower accuracy but greater methodological robustness. Another key distinction relates to model transparency: interpretability was limited in spectroscopic models using PCA/PLS-DA combinations, whereas the couple-based model allowed clearer assessment of predictor contributions. These findings underscore the need for transparent modeling strategies, standardized biomarker definitions, and rigorous validation to support clinical readiness.

Ethical, governance, and regulatory considerations were seldom reported in the included studies. None of the studies provided detailed descriptions of data governance practices, privacy protections, or assessments of potential bias and fairness—elements that are increasingly recognized as important for the safe deployment of AI/ML-driven decision-support tools. Contemporary regulatory and multi-agency guidance for AI/ML-enabled medical technologies places strong emphasis on lifecycle documentation, risk-based evaluation, and transparency to intended users, including clear communication of a model's intended use, limitations, and appropriate human oversight. In this context, explainability and user-facing transparency are not uniformly framed as fixed "requirements" across all use cases, but they are increasingly highlighted as key enablers of trust, reproducibility, and safe integration into clinical workflows. Therefore, future studies in unexplained infertility should report governance safeguards more explicitly (e.g., privacy and access controls, dataset provenance, and bias monitoring), alongside transparent model reporting and validation, to support responsible clinical translation.

### Strengths and Limitations

This systematic review has several strengths. It is among the first to synthesize machine learning-based predictive

models focusing on unexplained infertility outcomes. The application of the PROBAST tool for bias assessment is one of its strongest aspects. By addressing unexplained infertility—a critical but often overlooked issue in reproductive medicine—this review contributes valuable insights to the literature.

Nonetheless, several limitations exist. The small number of included studies restricts generalizability. There were notable differences across studies in sample sizes, data types, and ML algorithms used, which complicate comparisons. Future studies should employ multi-center datasets to enhance generalizability and reliability.

Finally, although biochemical and metabolic predictors were comprehensively addressed, psychological, environmental, and financial factors remain underrepresented in existing models.

Overall, the reviewed studies offer preliminary proof-of-concept evidence that ML-based models may capture biochemical and metabolic signatures relevant to unexplained infertility; however, the current literature remains methodologically heterogeneous and numerically limited, precluding firm clinical recommendations. Future research should prioritize multi-center datasets, standardized outcome definitions and reporting, and rigorous external (and preferably prospective) validation. In addition, incorporating broader clinical and biopsychosocial variables—including hormonal profiles and patient-reported measures—may improve model generalizability and support the development of clinically deployable AI tools for unexplained infertility.

### Conclusion

This systematic review highlights a major gap in the existing literature: despite the growing interest in artificial intelligence and machine learning in reproductive medicine, highly focused studies on unexplained infertility remain extremely limited. The sparseness of eligible studies underscores the need for more disease-specific, methodologically rigorous ML research to advance predictive modeling in this area.

Future studies incorporating ML or alternative AI techniques should aim to integrate the multidimensional factors involved in unexplained infertility for a more personalized approach. Additionally, the limited number of studies and the heterogeneity in datasets, algorithm types, and performance metrics prevented direct comparisons or meta-analysis.

**Ethics Committee Approval:** Not applicable. This study is a systematic review and meta-analysis based on previously published data and does not involve direct human or animal participation.

**Informed Consent:** As this study utilizes data from previously published studies, informed consent was obtained in the original investigations.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** None declared.

**Authorship Contributions:** Concept: TT, ÖAA; Design: TT, ÖAA; Supervision: TT, ÖAA; Resource: TT, ÖAA; Materials: TT, ÖAA; Data collection and/or processing: TT, ÖAA; Analysis and/or interpretation: TT, ÖAA; Literature review: TT, ÖAA; Writing: TT, ÖAA; Critical review: TT, ÖAA.

**Peer-review:** Double blind peer-reviewed.

## References

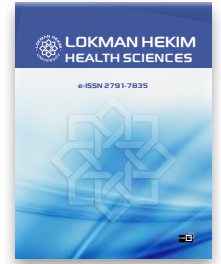
- Zegers-Hochschild F, Adamson GD, Dyer S, Racowsky C, de Mouzon J, Sokol R, et al. The International glossary on infertility and fertility care 2017. *Fertil Steril* 2017;108(3):393-406. [\[CrossRef\]](#)
- Raperport C, Desai J, Qureshi D, Rustin E, Balaji A, Chronopoulou E, et al. The definition of unexplained infertility: A systematic review. *BJOG* 2024;131(7):880-97. [\[CrossRef\]](#)
- Mahabbat NA, Khan TA, Elyas MF, Mahabbat AA, Oraif AM. Role of *in vitro* fertilization (ivf) in unexplained infertility management: a systematic review. *Cureus* 2024;16(10):e72527. [\[CrossRef\]](#)
- Niederberger C, Pellicer A, Cohen J, Gardner DK, Palermo GD, O'Neill CL, et al. Forty years of IVF. *Fertil Steril* 2018;110(2):185-324.e5.
- Griffin DK, Ogur C. Chromosomal analysis in IVF: just how useful is it? *Reproduction*. 2018;156(1):F29-F50. [\[CrossRef\]](#)
- Mol BW, Tjon-Kon-Fat R, Kamphuis E, van Wely M. Unexplained infertility: Is it over-diagnosed and over-treated? *Best Pract Res Clin Obstet Gynaecol* 2018;53:20-9. [\[CrossRef\]](#)
- Inhorn MC, Patrizio P. Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21<sup>st</sup> century. *Hum Reprod Update* 2015;21(4):411-26. [\[CrossRef\]](#)
- Jing X, Gu W, Zhang L, Miao R, Xu X, Wang M, et al. Coping strategies mediate the association between stigma and fertility quality of life in infertile women undergoing *in vitro* fertilization-embryo transfer. *BMC Womens Health* 2021;21(1):386. [\[CrossRef\]](#)
- Shingshetty L, Maheshwari A, McLernon DJ, Bhattacharya S. Should we adopt a prognosis-based approach to unexplained infertility? *Hum Reprod Open* 2022;2022(4):hoac046. [\[CrossRef\]](#)
- Dehghan S, Moghaddasi H, Rabiei R, Choobineh H, Maghooli K, Vahidi-Asl M. Machine learning in predicting infertility treatment success: A systematic literature review of techniques. *J Educ Health Promot* 2025;14:103. [\[CrossRef\]](#)
- Islam MN, Mustafina SN, Mahmud T, Khan NI. Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda. *BMC Pregnancy Childbirth* 2022;22(1):348. [\[CrossRef\]](#)
- Medenica S, Zivanovic D, Batkoska L, Marinelli S, Basile G, Perino A, et al. The future is coming: Artificial intelligence in the treatment of infertility could improve assisted reproduction outcomes-the value of regulatory frameworks. *Diagnostics* 2022;12:2979. [\[CrossRef\]](#)
- Giaxi P, Vivilaki V, Sarella A, Harizopoulou V, Gourounti K. Artificial intelligence and machine learning: an updated systematic review of their role in obstetrics and midwifery. *Cureus* 2025;17(3):e80394. [\[CrossRef\]](#)
- Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170:51-58. [\[CrossRef\]](#)
- Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015;4(1):1. [\[CrossRef\]](#)
- Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381. [\[CrossRef\]](#)
- Eldawlatly A, Alshehri H, Alqahtani A, Ahmad A, Al-Dammas F, Marzouk A. Appearance of population, intervention, comparison, and outcome as research question in the title of articles of three different anesthesia journals: a pilot study. *Saudi J Anaesth* 2018;12(2):283-6. [\[CrossRef\]](#)
- Depciuch J, Paja W, Pancercz K, Uzun Ö, Bulut H, Tarhan N, et al. Analysis of follicular fluid and serum markers of oxidative stress in women with unexplained infertility by Raman and machine learning methods. *J Raman Spectrosc* 2023;54:501511. [\[CrossRef\]](#)
- Jakubczyk P, Paja W, Pancercz K, Cebulski J, Depciuch J, Uzun Ö, et al. Determination of idiopathic female infertility from infrared spectra of follicle fluid combined with gonadotrophin levels, multivariate analysis and machine learning methods. *Photodiagnosis Photodyn Ther* 2022;38:102883. [\[CrossRef\]](#)
- Bachelot G, Lévy R, Bachelot A, Faure C, Czernichow S, Dupont C, et al. Proof of concept and development of a couple-based machine learning model to stratify infertile patients with idiopathic infertility. *Sci Rep* 2021;11(1):24003. [\[CrossRef\]](#)
- Nik Hazlina NH, Norhayati MN, Shaiful Bahari I, Nik Muhammad Arif NA. Worldwide prevalence, risk factors and psychological impact of infertility among women: a systematic review and meta-analysis. *BMJ Open* 2022;12(3):e057132. [\[CrossRef\]](#)
- Greil AL, Slauson-Blevins K, McQuillan J. The experience of infertility: A review of recent literature. *Sociol Health Illn* 2010;32:140-62. [\[CrossRef\]](#)
- Szkodziak F, Krzyżanowski J, Szkodziak P. Psychological

- aspects of infertility. A systematic review. *J Int Med Res* 2020;48(6):300060520932403. [\[CrossRef\]](#)
24. Duffy JMN, Adamson GD, Benson E, Bhattacharya S, Bhattacharya S, Bofill M, et al. Top 10 priorities for future infertility research: an international consensus development study. *Fertil Steril* 2021;115(1):180-90. [\[CrossRef\]](#)
25. Woods BM, Bray LA, Campbell S, Holland A, Mrug S, Ladores S. A review of the psychometric properties and implications for the use of the fertility quality of life tool. *Health Qual Life Outcomes* 2023;21(1):45. [\[CrossRef\]](#)
26. Chambers GM, Hoang VP, Sullivan EA, Chapman MG, Ishihara O, Zegers-Hochschild F, et al. The impact of consumer affordability on access to assisted reproductive technologies and embryo transfer practices: an international analysis. *Fertil Steril* 2014;101(1):191-8.e4. [\[CrossRef\]](#)
27. Dyer SJ, Patel M. The economic impact of infertility on women in developing countries a systematic review. *Facts Views Vis Obgyn* 2012;4(2):102-9.

<b>Appendix 1.</b> Full electronic search strategies for all databases				
Database	Date searched	Filters applied	Full Boolean search string	Records retrieved
PubMed	15–21 April 2025	2010–2025; Humans; English/Turkish	("Infertility"[MeSH] OR "infertility"[tiab] OR "unexplained infertility"[tiab] OR "idiopathic infertility"[tiab]) AND ("Machine Learning"[MeSH] OR "machine learning"[tiab] OR "artificial intelligence"[MeSH] OR "artificial intelligence"[tiab] OR "deep learning"[tiab] OR "neural network*" [tiab]) AND ("Predictive model*" [tiab] OR predict* [tiab] OR model* [tiab] OR prediction [tiab] OR classification [tiab])	n=203
Embase	15–21 April 2025	2010–2025; Humans; English/Turkish	('infertility'/exp OR infertilit* OR 'unexplained infertility' OR 'idiopathic infertility') AND ('machine learning'/exp OR 'machine learning' OR 'artificial intelligence'/exp OR 'artificial intelligence' OR 'deep learning' OR 'neural network*') AND (predict* OR 'predictive model*' OR model* OR classification OR algorithm*)	n=63
Scopus	15–21 April 2025	2010–2025; English/Turkish	TITLE-ABS-KEY(infertilit* OR "unexplained infertility" OR "idiopathic infertility" ) AND TITLE-ABS-KEY( "machine learning" OR "artificial intelligence" OR "deep learning" OR "neural network*" ) AND TITLE-ABS-KEY( predict* OR "predictive model*" OR model* OR classification OR algorithm* )	n=79
Web of Science	15–21 April 2025	2010–2025; English/Turkish	TS=(infertilit* OR "unexplained infertility" OR "idiopathic infertility") AND TS=("machine learning" OR "artificial intelligence" OR "deep learning" OR "neural network*") AND TS=(predict* OR "predictive model*" OR model* OR classification OR algorithm*)	n=157
Cochrane	15–21 April 2025	2010–2025; English/Turkish	("infertility" OR "unexplained infertility" OR "idiopathic infertility") AND ("machine learning" OR "artificial intelligence" OR "deep learning")	n=109
Google Scholar	15–21 April 2025	2010–2025; First 200 results screened	((Unexplained infertility [Title]) OR ((Infertility [Title]) AND ((Machine learning [Title]) AND ((Randomized [Title]) OR (Randomized Controlled Trial [Title]) OR (Randomized Clinical Trial [Title]))))))	n=301

**Appendix 2.** Studies excluded at full-text screening and reasons for exclusion

Study (author, year)	Full reference	Reason for exclusion
Liao et al., 2020	Liao S, Pan W, Dai WQ, Jin L, Huang G, Wang R, et al. <i>JAMA Network Open</i> . 2020;3(11):e2023654.	Heterogeneous infertility population; not specific to unexplained infertility
Liao et al., 2021	Liao S, Jin L, Dai WQ, Huang G, Pan W, Hu C, et al. <i>Int J Intelligent Systems</i> . 2021;36(3):1331–1344.	Heterogeneous infertility population; not specific to unexplained infertility
Balogun et al., 2018	Balogun JA, Egejuru NC, Idowu PA. <i>Computer Reviews Journal</i> . 2018;2(1):313–330.	Wrong population; general infertility prediction
Zhang et al., 2024	Zhang R, Zhou L, Hao X, Yang L, Ding L, Xing R, et al. <i>Metabolites</i> . 2024;14(9):492.	Wrong population; metabolic profiling in general infertility
Subha et al., 2024	Subha R, Nayana BR, Sumalatha P. <i>Eng Appl Artif Intell</i> . 2024;127:107400.	Wrong population; general infertility detection
Khan et al., 2024	Khan FM, Akhter MS, Khan IU, Haider ZA, Khan NH. <i>Int J Innovations Sci Technol</i> . 2024;6(2):943–960.	Wrong population; general infertility classification
Vats et al., 2022	Vats S, Sengupta A, Chaurasia A, Narad P. In: <i>Innovations in Computational Intelligence and Computer Vision</i> . 2022:547–555.	Wrong condition; PCOS-focused study
Raef & Ferdousi, 2019	Raef B, Ferdousi R. <i>Acta Informatica Medica</i> . 2019;27(3):205–210.	Review article; non-original research
Hassan et al., 2020	Hassan MR, Al-Insaf S, Hossain MI, Kamruzzaman J. <i>Neural Comput Appl</i> . 2020;32(7):2283–2297.	Wrong population; IVF pregnancy outcome prediction
Mehrjerd et al., 2022 (Sci Rep)	Mehrjerd A, Rezaei H, Eslami S, Ratna MB, Khadem Ghaebi N. <i>Scientific Reports</i> . 2022;12:7216.	Wrong population; general infertility treatment outcomes
Jha et al., 2024	Jha T, Sirisha M, Bhargavi MS. In: <i>2024 Int. Conf. for Women in Computing (InCoWoCo)</i> . IEEE; 2024:1–7.	Wrong condition; PCOS/PCOD diagnosis
Shofiyah & Mahmudy, 2023	Shofiyah S, Mahmudy WF. <i>8<sup>th</sup> Int. Conf. Sustainable Information Engineering &amp; Tech</i> . 2023:235–240.	Review article; non-original
Barnett-Itzhaki et al., 2020	Barnett-Itzhaki Z, Elbaz M, Buttermann R, Amar D, Amitay M, Racowsky C, et al. <i>J Assist Reprod Genet</i> . 2020;37(10):2405–2412.	Wrong population; general IVF population
Abdullah et al., 2023	Abdullah KAL, Atazhanova T, Chavez-Badiola A, Shivhare SB. <i>Reproductive Sciences</i> . 2023;30(4):1006–1016.	Review/conceptual article
Allameh et al., 2021	Allameh F, Fallah KM, Zadehmodarres S, Abedi AR, Eslami MJ, Hajian MR, et al. <i>Machine learning approaches to predict intrauterine insemination success rate</i> . 2021.	Wrong population; IUI prediction
Medenica et al., 2022	Medenica S, Zivanovic D, Batkoska L, Marinelli S, Basile G, Perino A, et al. <i>Diagnostics</i> . 2022;12(12):2979.	Review article; regulatory/ethical focus
Wang et al., 2024	Wang C, Johansson AL, Nyberg C, Pareek A, Almqvist C, Hernandez-Diaz S, Oberg AS. <i>Fertil Steril</i> . 2024;122(1):95–105.	Outcome not aligned; prediction of pregnancy-related complications
Mehrjerd et al., 2022 (Endometrial thickness)	Mehrjerd A, Rezaei H, Eslami S, Khadem Ghaebi N. In: <i>Challenges of Trustable AI and Added-Value on Health</i> . 2022:264–268.	Study used classical statistics only (no ML/AI)
Goyal et al., 2020	Goyal A, Kuchana M, Ayyagari KPR. <i>Scientific Reports</i> . 2020;10:20925.	Wrong population; general IVF live-birth prediction
Dehghan et al., 2024	Dehghan S, Rabiei R, Choobineh H, Maghooli K, Nazari M, Vahidi-Asl M. <i>PLoS One</i> . 2024;19(10):e0310829.	Wrong population; general IVF success prediction
Pouresmaeili et al., 2023	Pouresmaeili F, Alidoost S, Azimirad M, Azizmohammad Looha M, Emami Meibodi A, Abedin-Do A, et al. <i>Mol Biol Rep</i> . 2023;50(11):8785–8797.	Wrong population; unexplained recurrent miscarriage (RPL), not unexplained infertility
Koshy & Anuradha, 2022	Koshy S, Anuradha K. <i>Int J Eng Technol Manag Sci</i> . 2022;6:287.	Review article; not original ML/AI research
GhoshRoy et al., 2023	GhoshRoy, D., Alvi, P. A., & Santosh, K. C. (2023). <i>Journal of Medical Systems</i> . 47(1), 91.	<b>Review / non-original research; Wrong population</b>



## SYSTEMATIC REVIEW

# Artificial Intelligence in Neurosurgical Education: A Systematic Review of Technical Skills Training, Clinical Reasoning, and Surgical Planning

Ömer Selçuk Şahin, Samet Dinç

Department of Neurosurgery, Etlık City Hospital, Ankara, Türkiye

## Abstract

**Introduction:** Artificial intelligence (AI) and machine learning (ML) are increasingly used in neurosurgical education to mitigate limitations of apprenticeship-based training (restricted operative exposure, duty-hour constraints) and to enable objective, scalable competency assessment. This systematic review synthesized and critically appraised evidence on AI/ML applications for technical skills training, clinical reasoning support, and surgical planning.

**Methods:** Following Preferred Reporting Items for Systematic Reviews and Meta-analyses 2020, we searched seven databases (SciSpace Deep Review, SciSpace Basic Search, SciSpace Full-Text Search, Web of Science Core Collection, PubMed, Google Scholar, and arXiv) for English-language, peer-reviewed studies published January 2010–January 2026. Two reviewers independently screened records, extracted data, and assessed risk of bias using design-appropriate appraisal tools. Given methodological heterogeneity, a narrative synthesis was conducted.

**Results:** From 789 records, 36 studies met the inclusion criteria. Most focused on technical skills training (69.4%), followed by surgical planning (27.8%); fewer evaluated clinical reasoning support. AI-based assessment systems differentiated expertise with 83–100% accuracy. AI-augmented tutoring and feedback systems yielded improvements comparable to expert instruction (effect sizes 0.20–0.66). Common limitations included small sample sizes, single-center designs, and limited external validation.

**Discussion and Conclusion:** AI/ML technologies demonstrate clinically meaningful benefits for neurosurgical technical skills training. Cognitive and decision-support applications remain less mature and require multi-institutional validation, standardized outcomes, and longitudinal evaluation to support broader curricular integration.

**Keywords:** Artificial intelligence; Clinical decision-making; Competency-based medical education; Machine learning; Neurosurgical education; Simulation training

Neurosurgery is regarded as one of the most technically demanding and cognitively complex medical specialties. It requires advanced psychomotor coordination, refined visuospatial skills, and high-stakes

clinical decision-making under conditions of uncertainty. Neurosurgical education has long relied predominantly on apprenticeship-based training models, in which expertise is acquired through gradual clinical exposure under the

**Cite this article as:** Şahin ÖS, Dinç S. Artificial Intelligence in Neurosurgical Education: A Systematic Review of Technical Skills Training, Clinical Reasoning, and Surgical Planning. *Lokman Hekim Health Sci* 2026;6(2):303–312.

**Correspondence:** Ömer Selçuk Şahin, M.D. Etlık Şehir Hastanesi, Nöroşirürji Kliniği, Ankara, Türkiye

**E-mail:** oselcuks@gmail.com **Submitted:** 23.02.2026 **Revised:** 23.03.2026 **Accepted:** 31.03.2026 **Available Online:** 03.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



guidance of experienced professionals. Despite its efficacy, this paradigm is increasingly constrained by several factors. Chief among these are duty-hour regulations, heightened patient-safety expectations, growing procedural complexity, and variability in operative case exposure, particularly for rare or high-risk procedures.<sup>[1,2]</sup>

In response to these structural challenges, artificial intelligence (AI) and machine learning (ML) technologies have emerged as promising tools to augment neurosurgical education. AI systems can process high-dimensional data, identify latent performance patterns, and generate adaptive feedback at scale. Within the domain of neurosurgical training, the potential of AI applications has been thoroughly investigated across a wide range of educational settings. These applications encompass a diverse array of functions, including simulation-based technical skills training, automated performance assessment, intelligent tutoring systems, surgical planning support, and early cognitive decision-support tools.<sup>[1–35]</sup>

Simulation-based education has emerged as a pivotal component of contemporary neurosurgical training, offering controlled, standardized environments conducive to skill acquisition. Virtual reality (VR) and mixed reality (MR) platforms have demonstrated particular utility in neurosurgical education, enabling immersive anatomical visualization and procedural rehearsal in complex scenarios such as skull base tumor resection.<sup>[36–38]</sup> The integration of AI into these VR and simulation platforms has led to the rapid evolution of these tools from formative practice instruments to objective competency assessment instruments. A multitude of studies have demonstrated the capacity of ML-driven assessment systems to accurately differentiate levels of surgical expertise by meticulously analyzing kinematic, temporal, force-based, and video-derived metrics.<sup>[2,6–9,11,16,30]</sup> These systems offer advantages over traditional subjective rating scales in consistency, reproducibility, and granularity of feedback, while AI-augmented VR simulators provide standardized training experiences that address variability in clinical case exposure.<sup>[36,37]</sup>

Beyond episodic scoring, deep learning-based approaches enable continuous monitoring of surgical performance and analysis of the learning curve, allowing objective tracking of skill progression over time.<sup>[3,6,16]</sup> These competencies are closely aligned with competency-based medical education frameworks, which emphasize progression based on demonstrated performance rather than time-based criteria.<sup>[4,5]</sup> The findings of randomized controlled trials (RCTs) further indicate that AI-augmented intelligent tutoring

systems can achieve learning outcomes comparable to, and in some cases exceeding, those obtained through traditional expert instruction.<sup>[4,8,10,13]</sup>

Concurrently with the cultivation of technical proficiencies, there has been a marked surge in the exploration of AI applications in neurosurgical planning and decision support. ML models that incorporate imaging, demographic, and clinical variables have demonstrated greater predictive accuracy than clinician judgment alone, particularly when used as adjunctive tools.<sup>[18,21,26]</sup> From an educational perspective, exposure to such systems may introduce trainees to data-driven planning paradigms and support reflective comparison between human and algorithmic reasoning.

Despite rapid growth in the field, the extant literature on AI in neurosurgical education remains heterogeneous and underdeveloped. The extant literature focuses predominantly on technical skills training and assessment, while there is a paucity of research on applications that explicitly target clinical reasoning as an independent educational objective.<sup>[14,15,34]</sup> Furthermore, numerous studies are constrained by several factors, including limited sample sizes, single-center designs, reliance on simulated outcomes, and inadequate external validation of ML models.<sup>[6–9,14,17,34]</sup> The present systematic review addresses this gap by synthesizing and critically appraising evidence from 35 included studies on AI and ML applications in neurosurgical education.

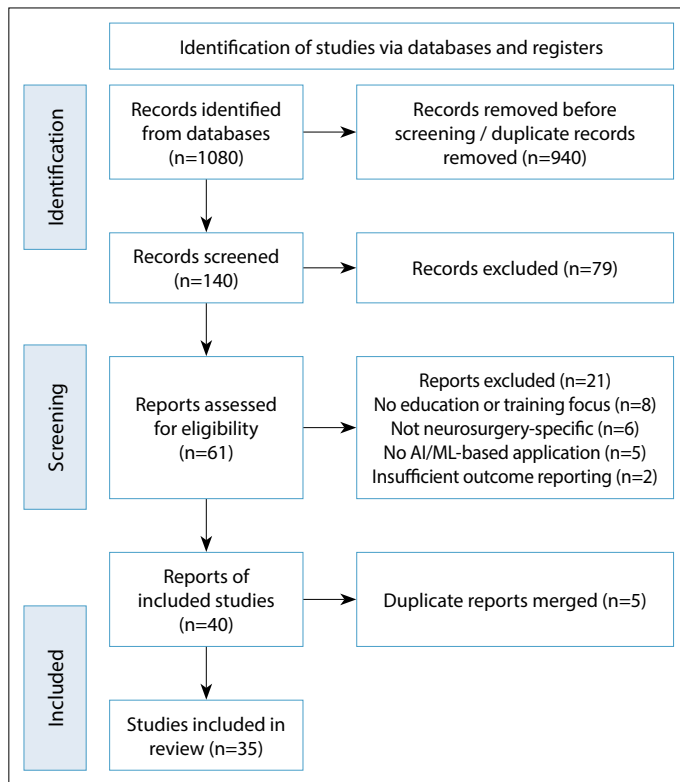
## Materials and Methods

### Study Design and Reporting Framework

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. A protocol was developed a priori to define objectives, eligibility criteria, information sources, study selection procedures, and the synthesis approach.<sup>[39]</sup> Ethical approval was deemed unnecessary in this instance; the scope was confined to analyzing previously published literature.

### Information Sources and Search Strategy

In January of 2026, a series of searches was conducted. The PubMed (MEDLINE) and Web of Science (WoS) Core Collection were identified as the primary databases for records reported in the PRISMA flow diagram. Concurrently, Google Scholar and SciSpace were used to enhance sensitivity and validate coverage.



**Figure 1.** PRISMA 2020 flow diagram for the study selection process. A total of 1080 records were identified through database searches, with no records identified from registers. After removal of 940 duplicate records, 140 records were screened by title and abstract, of which 79 were excluded. Full-text retrieval was sought for 61 reports, all of which were successfully retrieved and assessed for eligibility. Following full-text assessment, 21 reports were excluded with reasons, resulting in 35 studies included in the final qualitative synthesis.

Search strings combined controlled vocabulary (when available) and free-text terms across four concepts: The following areas of study are of particular interest: AI/ML, neurosurgery, education/training, populations, and simulation/VR/augmented reality (AR) or objective technical skill assessment. A variety of strategies were employed; each adapted to the specific indexing and syntax characteristics of the respective platform. The search was confined to the English-language publications from January 2010 to January 2026.

In WoS, the Advanced Search (TS) results were refined to include only article and review document types, excluding proceedings papers, yielding a total of 189 records. A MeSH and Title/Abstract hybrid strategy, restricted to journal articles and reviews, was utilized in PubMed, yielding 173 records. Due to the broad nature of Google Scholar and its lack of reproducibility, the scope of the results was constrained to the period from 2010 to 2026, with a sorting criterion based

on relevance. The screening process was further restricted to the initial 200 records from a total of 1,460 hits. A separate title-only (intitle): Google Scholar validation search identified 89 records, thereby confirming the scope of coverage. The use of SciSpace's full-text search and deep review was instrumental in expanding coverage and identifying existing lacunae. The records obtained from Google Scholar and SciSpace were included in the study only if they met the established eligibility criteria and were non-duplicate relative to the records identified through the primary databases. A comprehensive account of the search process is available in the supplementary appendix, specifically Appendix 1.

### Eligibility Criteria

Eligible studies were peer-reviewed journal articles or reviews involving medical students, neurosurgery residents/fellows, or practicing neurosurgeons that evaluated AI/ML applications with an explicit educational purpose in neurosurgery. Eligible domains included technical skills training, simulation-based education, VR/AR, and objective technical skill assessment.

Studies were excluded if they addressed clinical/diagnostic AI without an educational component, focused on non-neurosurgical specialties, did not report relevant educational or performance outcomes, or were editorials, conference abstracts, commentaries, or other non-peer-reviewed publications.

### Study Selection

All records were imported into a reference management system and deduplicated before screening. Two reviewers independently screened titles/abstracts, followed by full-text assessment of potentially eligible reports. Discrepancies were resolved by consensus. When multiple records referred to the same underlying study (e.g., a preprint and a journal version), the records were merged, and the most complete peer-reviewed version was retained. The selection process and the reasons for full-text exclusion are summarized in Figure 1.

### Data Extraction and Synthesis

Two reviewers independently extracted data using a standardized form capturing study design, participant characteristics, educational domain, AI/ML approach, comparator (if applicable), and reported educational or performance outcomes. Owing to heterogeneity in designs, interventions, and outcomes, findings were synthesized narratively.

## Risk of Bias (RoB) and Quality Appraisal

RoB and methodological quality were assessed using design-specific tools. These assessments were performed to contextualize and interpret the body of evidence and were not used to exclude studies.

Accordingly, RCTs were assessed using the Cochrane RoB 2 tool, non-randomized and quasi-experimental studies were evaluated using risk of bias in non-randomized studies of interventions, and predictive model studies were assessed using the prediction model risk of bias assessment tool (PROBAST).

## Results

### Study Selection

Searches across the electronic sources identified 1,080 records. After removal of 940 duplicate records, 140 records remained for title and abstract screening, of which 79 were excluded for failing to meet the predefined eligibility criteria. Sixty-one reports were sought for full-text retrieval; all were successfully retrieved and assessed for eligibility. Following full-text evaluation, 21 reports were excluded with documented reasons. Of the remaining 40 reports, 5 referred to duplicate publications of the same underlying study and were merged, yielding 35 unique studies included in the qualitative synthesis. The study selection process is summarized in Figure 1.

### Study Characteristics

The 35 studies included in this review were published between 2017 and 2025, with a clear inflection point after 2020, reflecting accelerating scholarly interest in AI-enabled neurosurgical education. Geographically, the evidence base was dominated by studies conducted in North America (51.4%), followed by Europe (22.9%), with additional contributions from Asia and multi-continental collaborative networks, underscoring the field's global expansion.

Methodologically, the included literature was diverse. RCTs accounted for 5 studies (14.3%), while quasi-experimental designs comprised 7 studies (20.0%). Observational cohort studies accounted for the most significant proportion (n=10, 28.6%), followed by cross-sectional studies (n=6, 17.1%). Evidence synthesis studies, including systematic and scoping reviews, accounted for 4 studies (11.4%), and pilot studies or case series accounted for the remaining 3 (8.6%).

Across primary empirical studies, sample sizes ranged from 14 to 156 participants, encompassing medical students, neurosurgery residents and fellows, and practicing

neurosurgeons. Collectively, these characteristics highlight both the methodological heterogeneity and the evolving maturity of the literature on AI-supported educational interventions in neurosurgery.

### Educational Focus Areas

Majority of the included studies focused on technical skills training (n=25, 71.4%), indicating the predominant use of AI and ML methods for performance measurement, automated feedback, and procedural proficiency in neurosurgical training. A smaller yet noteworthy proportion of studies focused on surgical planning training (n=10, 28.6%), predominantly using AI-assisted simulation environments and decision support interfaces.

A thorough review of the extant literature indicates that clinical reasoning has not been explored as a standalone educational domain. Conversely, reasoning-related constructions were implicitly incorporated within surgical planning tools or hybrid training frameworks. In these frameworks, cognitive processes were assessed indirectly alongside technical or procedural outcomes, rather than as independent learning endpoints.

### AI and ML Technologies

Conventional supervised ML classifiers, including support vector machines, k-nearest neighbors, naive Bayes, and random forests, were predominantly utilized for automated performance assessment (n=25, 71.4%). The implementation of deep learning architecture (n=15, 42.9%) facilitated continuous performance monitoring and video-based analysis. Computer vision systems were employed in 22.9% of the studies. A review of the extant literature reveals that fewer than one-third of the studies reported external validation

### Technical Skills Training

AI-based interventions for technical skills training demonstrated the strongest and most consistent evidence of efficacy. Automated assessment systems have shown the ability to accurately differentiate expertise levels across simulated tasks, with reported accuracies ranging from 83% to 100%.<sup>[2,6-9,11,16,30]</sup> Siyar et al.<sup>[2]</sup> achieved 91.7% accuracy using Fuzzy K-Nearest Neighbors. Winkler-Schwartz et al.<sup>[7]</sup> reported a 90% accuracy rate in distinguishing between four levels of expertise. Li et al.<sup>[9]</sup> achieved an accuracy of 92.41% with an area under the curve of 0.98. Karlik et al.<sup>[11]</sup> demonstrated a 100% classification accuracy using a hybrid fuzzy clustering neural network (Table 1).

**Table 1.** Distribution of AI/ML technologies in included studies (n=35)

AI/ML technology	n	%
Machine learning classifiers	25	69.4
Support vector machines	12	33.3
K-nearest neighbors	10	27.8
Naive bayes	8	22.2
Decision trees/Random forests	7	19.4
Deep learning/Neural networks	15	41.7
Convolutional neural networks	8	22.2
LSTM/Recurrent networks	5	13.9
Fully connected networks	6	16.7
Computer vision	8	22.2
Hybrid/Ensemble methods	6	16.7
Large language models	2	5.6

AI: Artificial intelligence; ML: Machine learning; LSTM: Long short-term memory.

A substantial body of research has emerged from RCTs, which have reported performance improvements comparable to those observed with traditional expert instruction through the implementation of AI-augmented intelligent tutoring systems.<sup>[4,8,10,13]</sup> As demonstrated by Fazlollahi et al.,<sup>[4]</sup> the virtual operative assistant has been shown to improve expertise scores by 0.66 points (95% confidence interval 0.55–0.77,  $p < 0.001$ ). Giglio et al.<sup>[8]</sup> demonstrated that AI-augmented instruction resulted in higher performance scores (mean difference 0.20,  $p = 0.02$ ). In a related study, Yilmaz et al.<sup>[13]</sup> showed that AI instruction led to a significant improvement in composite scores ( $p = 0.017$ ,  $p = 0.005$ ). In contrast, the administration of human instruction resulted in a decline in performance ( $p = 0.004$ ). As illustrated in Table 2, the study's primary outcomes are summarized.

### Surgical Planning and Clinical Reasoning

In a systematic/scoping review of AI applications in surgical planning, Senders et al.<sup>[20]</sup> reported a median improvement in accuracy of approximately 13% across included studies (as recalculated in the present review). In their systematic review of AI applications in surgical planning, Senders et al.<sup>[20]</sup> reported accuracy improvements; recalculating their data in the present review yielded a median improvement of approximately 13%.

From an educational perspective, the most significant benefits were observed when trainees engaged in reflective practice with AI outputs, utilizing them to compare algorithmic predictions with expert reasoning.<sup>[3,5,28]</sup> However, the findings of these studies were predominantly

short-term and simulation-based, with no study assessing clinical reasoning as an independent educational outcome. Consequently, the evidence for sustained improvement in higher-order cognitive skills remains inconclusive. To integrate the identified AI technologies, academic domains, and outcome dimensions across technical skills training, clinical reasoning, and surgical planning, a conceptual framework was developed to summarize AI's roles in neurosurgical education (Fig. 2).

The framework synthesizes evidence from the included studies, illustrating how AI technologies are applied across technical skills training, clinical reasoning, and surgical planning, and how these applications relate to assessment, feedback, and decision-support outcomes in neurosurgical education.

### Quality Assessment

Among the studies that employed PROBAST to assess prediction models (n=31), the most prevalent concerns were identified in the analysis domain (18/31, 58.1%) and the outcome domain (23/31, 74.2%), while participant selection was less frequently highlighted (11/31, 35.5%). The limitations of the studies included small sample sizes, single-center designs, reliance on simulated outcomes, and limited external validation. In PROBAST-based prediction models, the most common sources of bias were small training datasets (58.1%), lack of external validation (74.2%), and potential overfitting (35.5%). Despite the limitations, the uniformity of the findings lends credibility to the conclusions drawn.

### Discussion

This systematic review synthesizes evidence from 35 studies examining AI/ML applications with an explicit educational purpose in neurosurgery. A thorough review of the extant literature reveals an uneven evidence base across educational domains. The most robust and consistent support is for simulation-based technical skills training and objective performance assessment.<sup>[3,7,14,15,17]</sup> AI-enabled systems have demonstrated consistent proficiency in discriminating expertise levels and providing precise, replicable feedback across diverse platforms and tasks. This capability substantiates their function as scalable measurement instruments within competency-based training frameworks,<sup>[8,14,15,17]</sup> aligning with Miller's pyramid of clinical competence, which positions technical performance ("shows how") as a foundational prerequisite for autonomous clinical practice.<sup>[40]</sup> The observed accuracy

**Table 2.** Summary of key studies: AI technologies, applications, and outcomes

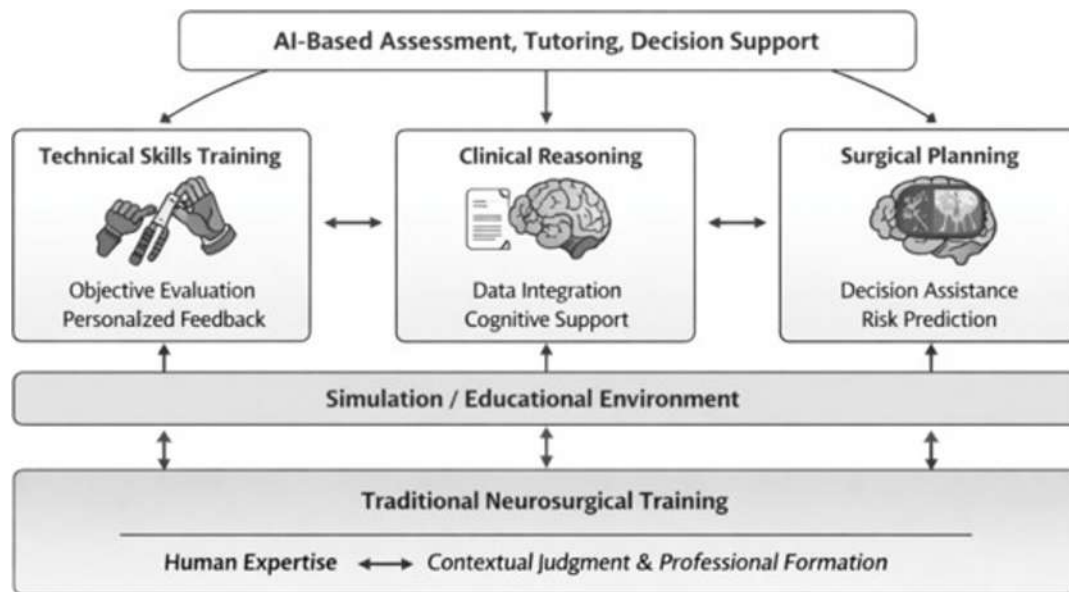
Study	AI technology	Educational application	Sample	Key outcomes
Siyar et al. <sup>[2]</sup>	Fuzzy KNN, SVM	VR tumor resection assessment	115	91.7% accuracy, 8.3% EER
Winkler-Schwartz et al. <sup>[7]</sup>	KNN, Naive Bayes, SVM	Expertise classification in VR	50	90% accuracy with KNN
Yilmaz et al. <sup>[6]</sup>	Deep neural network, LSTM	Continuous bimanual monitoring	50	4 expertise levels (p<0.001), R <sup>2</sup> =27.7%
Fazlollahi et al. <sup>[4]</sup>	Deep learning (ICEMS)	AI tutoring versus expert (RCT)	70	VOA improved 0.66 pts (p<0.001)
Giglio et al. <sup>[8]</sup>	AI tutoring system	AI-augmented instruction (RCT)	87	Mean difference 0.20 (p=0.02)
Yilmaz et al. <sup>[13]</sup>	AI intelligent instruction	Real-time AI versus in-person (RCT)	25	AI improved (p=0.017); human decreased
Li et al. <sup>[9]</sup>	SVM with shapley values	Personalized VR assessment	79	92.41% accuracy, AUC=0.98
Karlik et al. <sup>[11]</sup>	Fuzzy clustering NN	Expertise classification in VR	NS	100% accuracy with FCNN
Singh et al. <sup>[15]</sup>	Naive Bayes, SVM, DT	Craniotomy drilling assessment	22	90.0% accuracy
Ledwos et al. <sup>[16]</sup>	KNN for metric selection	Learning curve analysis	50	5/6 metrics significant (p<0.05)
Senders et al. <sup>[20]</sup>	Various ML models	Surgical planning (SR)	23 studies	13% median accuracy improvement
Sugiyama et al. <sup>[12]</sup>	Semantic segmentation	Microsurgical video analysis	14	Strong correlation with ratings
Witten et al. <sup>[32]</sup>	ML segmentation	Neuroanatomical classification	NS	91.8% test accuracy
Reich et al. <sup>[5]</sup>	Artificial neural networks	Competency-based VR training	NS	Objective tracking of skill progression
Mirchi et al. <sup>[29]</sup>	Explainable AI	Simulation-based training (VOA)	NS	Transparent and personalized feedback
Pangal et al. <sup>[30]</sup>	Video-based ML	Automated performance metrics	NS	Automated expertise differentiation
Bocanegra-Becerra <sup>[18]</sup>	Various ML algorithms	Pre-operative planning (SR)	NS	Integration of imaging/clinical data

NS: Not specified; EER: Equal error rate; SR: Systematic/Scoping review; DT: Decision tree; NN: Neural network; VR: Virtual reality; RCT: Randomized controlled trial; FCNN: Fuzzy clustering neural network; VOA: Virtual operative assistant; AI: Artificial intelligence; SVM: Support vector machines; KNN: K-nearest neighbors; ICEMS: Intelligent continuous expertise monitoring system; AUC: Area under the curve.

of AI systems in expertise classification (83–100%) suggests their potential utility in systematically documenting learner progression through the stages of skill acquisition described in the Dreyfus model—from novice reliance on explicit rules to expert intuitive performance.<sup>[41]</sup>

A substantial body of research, supported by empirical evidence from randomized and comparative studies, suggests that AI-augmented instruction, particularly when integrated within structured simulation curricula, can enhance technical performance and learning efficiency.<sup>[4,5]</sup> The merits of AI in education are most clearly delineated when it functions as an augmentative mechanism, a standardizing agent of feedback, and an enabler of deliberate practice as conceptualized by Ericsson.<sup>[42]</sup> Deliberate practice, characterized by focused repetition with immediate corrective feedback, represents the cornerstone of expert performance development. AI systems that provide granular, real-time performance metrics and individualized coaching operationalize this framework at scale, enabling trainees to engage in high-repetition practice with consistent feedback quality independent of faculty availability. While virtual and MR platforms provide immersive training environments,<sup>[38]</sup> AI integration adds objective, quantitative assessment capabilities that VR alone cannot deliver, exemplifying how complementary technologies address different dimensions of surgical competence. In such cases, AI serves to complement, rather than supplant, expert supervision. This approach preserves the educational value of mentorship while concomitantly reducing variability in assessment.<sup>[3–5,7]</sup>

Conversely, applications designed for surgical planning education were less numerous and more heterogeneous. Despite the demonstrated potential advantages of AI-assisted planning tools when used in conjunction with human judgment, educational outcomes were frequently inferred indirectly rather than measured longitudinally.<sup>[1,13]</sup> It is important to note that clinical reasoning was rarely operationalized as a standalone educational endpoint. Instead, reasoning-related elements were embedded within hybrid technical–cognitive or planning frameworks. This finding underscores the intricacies of measuring higher-order cognition and the limitations of current methodological approaches.<sup>[1,13]</sup> Within Bloom's taxonomy of educational objectives,<sup>[43]</sup> most AI applications in this review address lower cognitive domains (knowledge, comprehension, application) rather than higher-order processes such as analysis, synthesis, and evaluation—the very competencies that distinguish expert clinical reasoning. The predominant focus on



**Figure 2.** Conceptual framework of artificial intelligence applications in neurosurgical education.

psychomotor skill assessment reflects a natural alignment between simulation-generated kinematic data and ML algorithms, yet leaves unaddressed the critical cognitive skills that enable surgeons to adapt procedural knowledge to novel clinical contexts, anticipate complications, and make sound judgments under uncertainty.

This “cognitive lacuna” in the literature underscores a salient risk of automation bias, in which trainees may become excessively reliant on algorithmic outputs, potentially at the expense of developing their own clinical intuition. As the field of AI matures, it is imperative to transition from using AI as a passive scoring tool to a “cognitive partner.” This necessitates the incorporation of explainable AI (XAI) frameworks<sup>[29]</sup> that do not merely provide a performance score but also deconstruct the underlying cognitive architecture of surgical decision-making. Moreover, while AI systems demonstrate accuracy ranging from 83% to 100%, the absence of external validation in 74.2% of studies<sup>[15,31]</sup> constitutes a significant impediment to clinical translation. Future research must prioritize multi-institutional validation and the tracking of skill retention and transfer from simulated environments to the operating room. The actual value of AI in neurosurgery will not lie in its capacity to supplant the mentor; instead, it will be found in its ability to deconstruct the “black box” of surgical expertise, thereby providing a transparent, equitable, and evidence-based foundation for the next generation of mastery.<sup>[32,35]</sup>

In the context of surgical planning education, pedagogical risks merit explicit attention. When trainees interact with high-performing planning or prediction tools, they may develop overconfidence or a false sense of security, defer

prematurely to algorithmic recommendations, or reduce reflective comparison with expert reasoning, especially when model uncertainty and failure modes are not transparent. Such effects can inadvertently weaken deliberate practice in clinical judgment and promote “deskilling” in scenario appraisal. Mitigation strategies include supervised use with structured debriefing, calibrated trust training (requiring trainees to justify decisions independent of the model), presentation of confidence/uncertainty estimates, and adoption of explainable AI interfaces that highlight rationale and limitations rather than only outputs.

From a theoretical standpoint, the prevalence of technical skills studies in this field indicates the inherent congruence between simulation-generated data streams and ML methodologies. High-frequency kinematic and performance data are compatible with supervised learning approaches, thereby facilitating model development and validation.<sup>[8,14,15]</sup> Conversely, reasoning and judgment are context-dependent and less directly observable. This observation underscores the need for greater integration of educational theory, cognitive science, and explainable AI to ensure that AI outputs support learning rather than automation bias.<sup>[13]</sup>

This asymmetry in AI application, robust support for technical skills, but limited engagement with clinical reasoning, carries important educational implications. Competency in neurosurgery requires not only procedural fluency but also the cultivation of what Schön termed “reflection-in-action”:<sup>[44]</sup> The capacity to think critically during performance, recognize anomalies, and adaptively modify one’s approach. Current

AI systems excel at evaluating what trainees do (observable behaviors, tool paths, efficiency metrics) but cannot yet adequately assess how they think or why they make specific decisions. This limitation becomes particularly salient when considering the progression from novice to expert as described by the dreyfus model,<sup>[41]</sup> where advanced practitioners move beyond rule-following to develop holistic, context-sensitive judgment. Future AI applications must therefore evolve beyond passive performance measurement to actively scaffold metacognitive processes, prompting trainees to articulate their reasoning, compare their decision-making with expert approaches, and develop the reflective habits that characterize surgical expertise.

Several practical considerations must be considered. AI-enabled simulation has the potential to enhance practice density and facilitate individualized learning trajectories, while operating within constraints such as reduced operative exposure and limited faculty time.<sup>[3,7]</sup> However, the comprehensive implementation of the curriculum is contingent upon substantiating its correlation with clinical performance, skill retention, and program-level impact.<sup>[3,7,8,13]</sup> Equity considerations underscore the importance of multi-institutional validation and transparent reporting to mitigate algorithmic bias and site-specific overfitting.<sup>[3,13,17]</sup>

A review of the extant literature reveals several limitations. Firstly, the samples are often small, and studies are predominantly conducted at a single center. Furthermore, there is considerable methodological heterogeneity. These factors prevented conducting a quantitative synthesis. The application of external validation and standardized reporting remains constrained, thereby limiting the generalizability of findings.<sup>[3,13,17]</sup> At the review level, the available literature that has undergone peer review, as well as the variability across supplementary search platforms, may have influenced the extent of coverage.<sup>[3]</sup>

## Conclusion

A thorough review of the extant literature indicates that AI and ML applications have demonstrated high precision and reproducibility in simulation-based technical skills assessment. These technologies have proven effective as objective measurement instruments, offering standardized evaluation and scalable learning support that transcends the limitations of traditional apprenticeship models. The most significant instructional gains are derived from hybrid training frameworks where AI-augmented systems complement expert supervision by standardizing feedback and extending deliberate practice, rather than replacing the essential role of human mentorship.<sup>[28,30]</sup>

Conversely, while AI-assisted tools demonstrate potential in surgical planning, their application in cultivating and assessing clinical reasoning remains nascent and inconsistently operationalized. The assessment of learning outcomes in these higher-order cognitive domains is often based on inferences rather than on longitudinal, evidence-based metrics. The development of a cognitive foundation necessitates a transformation in educational paradigms, encompassing the integration of transparent, XAI frameworks. These frameworks should prioritize critical judgment over automation bias, thereby facilitating the advancement of knowledge and skills in a systematic and well-informed manner.<sup>[29]</sup>

In addition, the incorporation of ML algorithms for automated segmentation of operative neuroanatomy has demonstrated considerable potential to enhance the precision of preoperative planning. For instance, Witten et al.<sup>[32]</sup> demonstrated that AI-driven image segmentation can effectively delineate complex neuroanatomical structures, thereby providing trainees with a high-fidelity visual roadmap that facilitates better spatial orientation during the transition from simulation to the operating theater. The ultimate objective for neurosurgical training is to ensure that these innovations translate into improvements that are both generalizable and evidence-based.<sup>[17]</sup>

In contemplating the imminent integration of AI-driven innovations within clinical practice, it is imperative to address the prevailing methodological limitations. Future research must prioritize multicenter validation, tracking skill retention, and objective measurement of technical transfer to the operating room.<sup>[15,31]</sup> Adherence to consistent reporting standards and a focus on equitable, data-driven advancements are pivotal for the potential of AI to fundamentally redefine neurosurgical mastery and ensure superior patient-safety outcomes in an increasingly complex operative landscape.<sup>[32,35]</sup>

**Ethics Committee Approval:** Ethics committee approval was not required for this study because it is a systematic review and does not involve direct human or animal participants, patient intervention, or access to identifiable personal data.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The authors declared that this study received no financial support.

**Use of AI for Writing Assistance:** The authors would like to acknowledge the use of AI-assisted writing tools for language editing, structural refinement, and reference formatting. All intellectual content, including the study rationale, analysis, interpretation, and conclusions, is the original work of the authors, who take full responsibility for the manuscript.

**Authorship Contributions:** Concept: ÖSS; Design: ÖSS; Supervision: ÖSS, SD; Resource: ÖSS; Materials: ÖSS; Data collection and/or processing: ÖSS, SD; Analysis and/or interpretation: ÖSS, SD; Literature review: ÖSS; Writing: ÖSS, SD; Critical review: ÖSS.

**Peer-review:** Double blind peer-reviewed.

## References

1. Wu J, Liang X, Bai XF, Chen Z. SurgBox: agent-driven operating room sandbox with surgery copilot. *IEEE Int Conf Big Data* 2024;1449-58. [\[CrossRef\]](#)
2. Siyar S, Azarnoush H, Rashidi S, Winkler-Schwartz A, Bissonnette V, Ponnudurai N, et al. Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task. *Med Biol Eng Comput* 2020;58(6):1357-67. [\[CrossRef\]](#)
3. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Ledwos N, Del Maestro RF. Development and predictive validation of an intelligent, continuous assessment system for bimanual surgical skills. *Can J Neurol Sci* 2022;49(6):1-10. [\[CrossRef\]](#)
4. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. *JAMA Netw Open* 2022;5(2):e2149008. [\[CrossRef\]](#)
5. Reich A, Mirchi N, Yilmaz R, Ledwos N, Bissonnette V, Tran DH, et al. Artificial neural network approach to competency-based training using a virtual reality neurosurgical simulation. *Oper Neurosurg* 2022;23(1):31-9. [\[CrossRef\]](#)
6. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Christie S, Tran DH, et al. Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. *NPJ Digit Med* 2022;5(1):54. [\[CrossRef\]](#)
7. Winkler-Schwartz A, Yilmaz R, Mirchi N, Bissonnette V, Ledwos N, Siyar S, et al. Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. *JAMA Netw Open* 2019;2(8):e198363. [\[CrossRef\]](#)
8. Giglio B, Albeloushi A, Alhaj A, Alhantoobi M, Saeedi R, Davidovic V, et al. Artificial intelligence-augmented human instruction and surgical simulation performance. *JAMA Surg* 2025;160(4):1-8. [\[CrossRef\]](#)
9. Li F, Qin Z, Qian K, Liang S, Li C, Tai Y. Personalized assessment and training of neurosurgical skills in virtual reality: an interpretable machine learning approach. *Virtual Reality Intell Hardware* 2024;6(1):27-43. [\[CrossRef\]](#)
10. Piñera-Castro HJ, Borges-García C. Applications of artificial intelligence in neurosurgical education: a scoping review. *Egypt J Neurol Psychiatr Neurosurg* 2025;61(1):15. [\[CrossRef\]](#)
11. Karlik B, Yilmaz R, Winkler-Schwartz A, Mirchi N, Bissonnette V, Ledwos N, et al. Assessment of surgical expertise in virtual reality simulation by hybrid deep neural network algorithms. *Int J Artif Intell Expert Syst* 2021;10(3):47-59.
12. Sugiyama T, Tang M, Sugimori H, Sakamoto M, Fujimura M. Artificial intelligence-integrated video analysis of vessel area changes and instrument motion for microsurgical skill assessment. *Sci Rep* 2025;15(1):27898. [\[CrossRef\]](#)
13. Yilmaz R, Bakhaidar M, Alsayegh A, Abou Hamdan N, Fazlollahi AM, Tee T, et al. Real-time multifaceted artificial intelligence vs in-person instruction in teaching surgical technical skills: a randomized controlled trial. *Sci Rep* 2024;14(1):15130. [\[CrossRef\]](#)
14. Titov O, Bykanov A, Pitskhelauri D. Neurosurgical skills analysis by machine learning models: systematic review. *Neurosurg Rev* 2023;46(1):121. [\[CrossRef\]](#)
15. Singh R, Godiyal AK, Suri A. Craniotomy simulator with force myography and machine learning-based skills assessment. *Bioengineering* 2023;10(4):465. [\[CrossRef\]](#)
16. Ledwos N, Mirchi N, Yilmaz R, Winkler-Schwartz A, Sawani A, Fazlollahi AM, et al. Assessment of learning curves on a simulated neurosurgical task using metrics selected by artificial intelligence. *J Neurosurg* 2022;137(4):1160-71. [\[CrossRef\]](#)
17. Davids J, Manivannan S, Darzi A, Giannarou S, Ashrafian H, Marcus HJ. Simulation for skills training in neurosurgery: a systematic review, meta-analysis, and analysis of progressive scholarly acceptance. *Neurosurg Rev* 2021;44(4):1853-67. [\[CrossRef\]](#)
18. Bocanegra-Becerra JE, Neves Ferreira JS, Simoni G, Hong A, Rios-Garcia W, Eraghi MM, et al. Machine learning algorithms for neurosurgical preoperative planning: a scoping review. *World Neurosurg* 2025;194:123465. [\[CrossRef\]](#)
19. Danushka N, Wijesinghe D, Jayasinghe R, Attanayake D. AI-powered precision: transforming neurosurgical practice through intelligent technologies. In: *Artificial Intelligence*. London (UK): IntechOpen; 2025. [\[CrossRef\]](#)
20. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, et al. Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery* 2018;83(2):181-92. [\[CrossRef\]](#)
21. Dundar TT, Yurtsever I, Pehlivanoglu MK, Yildiz U, Eker A, Demir MA, et al. Machine learning-based surgical planning for neurosurgery: artificial intelligent approaches to the cranium. *Front Surg* 2022;9:863633. [\[CrossRef\]](#)
22. Khizar A. Artificial intelligence and neurosurgery: a revolution in the field. *Pak J Neurol Sci* 2024;18(4):244. [\[CrossRef\]](#)
23. Mofatteh M. Neurosurgery and artificial intelligence. *AIMS Neurosci* 2021;8(4):477-95. [\[CrossRef\]](#)
24. Awuah WA, Adebusoye FT, Wellington J, David L, Salam A, Weng Yee AL, et al. Recent outcomes and challenges of artificial intelligence, machine learning, and deep learning in neurosurgery. *World Neurosurg X* 2024;23:100301. [\[CrossRef\]](#)
25. Singh C, Gharde P, Verma P, et al. Artificial intelligence in neurosurgery: enhancing diagnosis, treatment and patient outcomes: a narrative review. *J Clin Diagn Res* 2025;19(9):PE01-PE05. [\[CrossRef\]](#)
26. Yin S, Ming J, Chen H, Sun Y, Jiang C. Integrating deep learning and radiomics for preoperative glioma grading using multi-center MRI data. *Sci Rep* 2025;15(1):36756. [\[CrossRef\]](#)

27. Maghrabi Y, Jamjoom AB, Algahtani A, Alshareef OH, Jamjoom OM, Alzahrani M. Highly cited artificial intelligence research studies published in neurosurgical journals: a bibliometric analysis. *Cureus* 2025;17(11):e98191. [\[CrossRef\]](#)
28. Sugiyama T, Sugimori H, Tang M, Fujimura M. Artificial intelligence for patient safety and surgical education in neurosurgery. *JMA J* 2025;8(1):76-85. [\[CrossRef\]](#)
29. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, DelMaestro RF. The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One* 2020;15(2):e0229596. [\[CrossRef\]](#)
30. Pangal DJ, Kugener G, Cardinal T, Lechtholz-Zey E, Collet C, Lasky S, et al. Use of surgical video-based automated performance metrics to predict blood loss and success of simulated vascular injury control in neurosurgery: a pilot study. *J Neurosurg* 2021;137(3):840-9. [\[CrossRef\]](#)
31. Escobar-Castillejos D, Barrera-Animas AY, Noguez J, Magana AJ, Benes B. Transforming surgical training with AI techniques for training, assessment, and evaluation: scoping review. *J Med Internet Res* 2025;27:e58966. [\[CrossRef\]](#)
32. Witten AJ, Patel NB, Cohen-Gadol AA. Image segmentation of operative neuroanatomy using machine learning. *Oper Neurosurg* 2022;23(4):e322-8. [\[CrossRef\]](#)
33. Konakondla S, Fong R, Schirmer CM. Simulation training in neurosurgery: advances in education and practice. *Adv Med Educ Pract* 2017;8:465-73. [\[CrossRef\]](#)
34. Harley JM, Tawakol T, Azher S, Quaiattini A, Maestro RD. The role of AI in neurosurgical education: an umbrella review. *Glob Surg Educ* 2024;3(1):83. [\[CrossRef\]](#)
35. Tariciotti L, Palmisciano P, Giordano M, Remoli G, Lacorte E, Bertani G, et al. Artificial intelligence-enhanced intraoperative neurosurgical workflow: current knowledge and future perspectives. *J Neurosurg Sci* 2022;66:139-50. [\[CrossRef\]](#)
36. Shao X, Yuan Q, Qian D, Ye Z, Chen G, Zhuang K, et al. Virtual reality technology for teaching neurosurgery of skull base tumor. *BMC Med Educ* 2020;20(1):3. [\[CrossRef\]](#)
37. Jain S, Timofeev I, Kirollos RW, Helmy A. Use of mixed reality in neurosurgery training: a single centre experience. *World Neurosurg* 2023;176:e68-76. [\[CrossRef\]](#)
38. Silvero Isidre A, Friederichs H, Mütter M, Gallus M, Stummer W, Holling M. Mixed reality as a teaching tool for medical students in neurosurgery. *Medicina (Kaunas)* 2023;59(10):1720. [\[CrossRef\]](#)
39. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. [\[CrossRef\]](#)
40. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65(9 Suppl):S63-7. [\[CrossRef\]](#)
41. Dreyfus SE, Dreyfus HL. A five-stage model of the mental activities involved in directed skill acquisition. Berkeley: University of California, Operations Research Center; 1980. [\[CrossRef\]](#)
42. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. *Acad Emerg Med* 2008;15(11):988-94. [\[CrossRef\]](#)
43. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of educational objectives: the classification of educational goals. Handbook I: cognitive domain. New York: David McKay Company; 1956.
44. Schön DA. The reflective practitioner: how professionals think in action. New York: Basic Books; 1983.

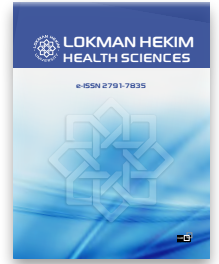
**Appendix 1.** Database-specific search strategies and retrieval details

This appendix reports the full search strategies, platform-specific adaptations, and retrieval decisions applied in this systematic review, ensuring transparency and reproducibility in accordance with PRISMA 2020 guidelines.

Table A1. Harmonized search strategy across databases

Database / platform	Search strategy (final, high-specificity)	Search date	Platform mode	Filters / restrictions applied	Hits retrieved (n)	Role in review
PubMed (MEDLINE)	MeSH + Title/Abstract hybrid query requiring AI/ML, neurosurgery, education/training, and simulation/VR or objective skill assessment.	January 2026	Advanced Search	English; Publication date 2010–Jan 2026; Journal Article OR Review	173	Primary database for systematic identification
Web of Science Core Collection	TS-based query requiring AI/ML, neurosurgery, education/training, and simulation/VR or skill assessment.	January 2026	Advanced Search (TS)	Article (and Review); Proceedings Papers excluded; English; 2010–2026	189	Primary database for systematic identification
Google Scholar	Keyword-based Boolean query aligned with the four-domain structure (AI/ML + neurosurgery + education + simulation/skills).	January 2026	Relevance-sorted search	Custom year range 2010–2026; first 200 results screened; peer-reviewed journal articles/reviews retained	1460	Supplementary source to enhance sensitivity
Google Scholar (title-only validation search)	intitle:(simulation OR "virtual reality" OR "skill assessment") AND (neurosurgery OR neurosurgical OR "spine surgery") AND ("artificial intelligence" OR "machine learning" OR "deep learning")	January 2026	intitle: operator	English; 2010–2026	89	Validation and specificity check
SciSpace – Full-Text Search	Full-text keyword query using the same four-domain structure (AI/ML, neurosurgery, education, simulation/skills).	January 2026	Full-Text Search	Year ≥2010; English where available; peer-reviewed focus during screening	100	Supplementary source for coverage enhancement
SciSpace – Deep Review	Structured AI-assisted query restricted to peer-reviewed AI/ML educational applications in neurosurgery involving simulation/VR or skill assessment.	January 2026	Deep Review (AI-assisted)	English; 2010–Jan 2026; peer-reviewed focus	529	Curated cross-check and gap detection

PRISMA Compliance and Reporting Notes: Search strategies were prospectively defined and consistently structured across platforms around four core concepts: artificial intelligence/machine learning, neurosurgery, education or training, and simulation-based or objective skill assessment outcomes. Database-specific syntax and filters were applied as required. Google Scholar searches were capped to ensure feasibility and reproducibility, and title-only searches were used solely as validation exercises. Retrieval counts were recorded prior to de-duplication and are summarized in the PRISMA 2020 flow diagram.



# Mapping Artificial Intelligence-based Assessment Domains in Pediatric Occupational Therapy: A Scoping Review

Zeynep Kolit, Duygu Mine Alataş, Büşra Elif Ayber

Department of Occupational Therapy, Lokman Hekim University Faculty of Health Sciences, Ankara, Türkiye

## Abstract

**Introduction:** Although artificial intelligence (AI) has increasingly been integrated into health and rehabilitation sciences, the assessment domains in which AI-based approaches are applied in pediatric occupational therapy have not yet been comprehensively mapped in the literature. The aim of this study was to systematically identify the assessment domains in which AI-based approaches have been applied in pediatric occupational therapy and to describe how these applications have been reported in the existing literature.

**Methods:** A scoping review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews. Searches were performed in CINAHL, OTseeker, PubMed/MEDLINE, Scopus, and Web of Science for English-language, peer-reviewed studies published between 2015 and 2025. Empirical studies reporting the use of AI-based approaches for assessment purposes in pediatric occupational therapy were included. Data were charted and synthesized descriptively to provide an overview of research characteristics and thematic patterns.

**Results:** Sixteen studies met the inclusion criteria. Most studies employed observational or validation-based designs and primarily utilized machine learning approaches, including deep learning and computer vision techniques. AI-based assessment applications were predominantly focused on motor and sensory-perceptual domains. Fewer studies addressed cognitive functioning and activities of daily living, while no studies explicitly examined psychosocial, environmental, or participation-related assessment domains. Image- and video-based data were the most frequently used modalities.

**Discussion and Conclusion:** The findings indicate that AI-based assessment research in pediatric occupational therapy has largely concentrated on performance-oriented domains, particularly motor and sensory-perceptual functioning. Important occupational therapy domains, such as participation, environmental context, and psychosocial functioning, remain underrepresented. This scoping review provides an overview of current research trends and highlights key gaps, offering a foundation to guide future interdisciplinary research and the development of more holistic, occupation-centered AI-based assessment approaches within health sciences.

**Keywords:** Artificial intelligence; Machine learning; Occupational therapy; Outcome assessment; Pediatrics

Artificial intelligence (AI) refers to a broad class of technological systems that enable machines to perform tasks typically requiring human intelligence, such as learning from data, recognizing complex patterns, and

making decisions. Over the past decade, AI has driven substantial advances in both basic and clinical research across multiple fields related to human health and behavior, including medicine, psychology, and education.<sup>[1,2]</sup> In the

**Cite this article as:** Kolit Z, Alataş DM, Ayber BE. Mapping Artificial Intelligence-based Assessment Domains in Pediatric Occupational Therapy: A Scoping Review. Lokman Hekim Health Sci 2026;6(2):313–322.

**Correspondence:** Zeynep Kolit, M.D. Lokman Hekim Üniversitesi, Sağlık Bilimleri Fakültesi, Ergoterapi Anabilim Dalı, Ankara, Türkiye

**E-mail:** zeynep.kolit@lokmanhekim.edu.tr **Submitted:** 04.03.2026 **Revised:** 29.04.2026 **Accepted:** 08.05.2026 **Available Online:** 11.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



fields of health and rehabilitation, AI plays an innovative role in enhancing the quality of healthcare services through applications such as clinical decision support systems, the development of intervention strategies, improved accuracy and efficiency of assessments, and more precise data analysis.<sup>[3,4]</sup> Although AI is used for various purposes across multiple fields, evidence from the literature indicates that its application in occupational therapy has primarily focused on functional assessments, personalized rehabilitation programs, and intervention planning.<sup>[4,5]</sup>

In pediatric occupational therapy, assessment is a fundamental process that enables the holistic identification of the physical, cognitive, sensory-perceptual, and psychosocial challenges children encounter during activities of daily living (ADL).<sup>[6]</sup> Through this comprehensive assessment process, occupational therapists are able to support goal setting and the structuring of interventions by simultaneously considering the child's individual needs, environmental factors, and activity demands.<sup>[5]</sup> Occupational therapy assessments largely rely on therapists' clinical experience and observational judgment, which may lead to inconsistencies among evaluators.<sup>[7]</sup> In addition, the time-consuming nature of the assessment process may increase therapists' workload while causing fatigue and reduced motivation in children, thereby negatively affecting assessment accuracy.<sup>[8]</sup>

In recent years, AI-based assessment approaches have increasingly been adopted to address the limitations and challenges of conventional evaluation methods.<sup>[9]</sup> These approaches offer therapists objective, standardized, and data-driven information to enhance the assessment process and can automate data analysis, reducing the time and effort required for manual evaluation and interpretation.<sup>[4,8]</sup> In this context, despite the growing number of AI-driven applications in the health sciences, studies that systematically and comprehensively examine the assessment domains in which these technologies are applied in pediatric occupational therapy, as well as the types of data and AI approaches employed, remain limited.<sup>[5,9]</sup> Therefore, a scoping review was deemed necessary to comprehensively map the existing literature. This scoping review aims to elucidate how AI-based applications are positioned across assessment domains in pediatric occupational therapy, including sensory-perceptual, motor, cognitive, psychosocial, environmental, ADL, and participation domains; to classify the technologies and AI methods utilized; and to identify gaps in the current literature, thereby providing a guiding framework for future research.

## Materials and Methods

### Study Design

This study was conducted as a scoping review with the aim of systematically identifying the assessment domains in which AI has been used in pediatric occupational therapy and describing how these applications have been reported in the literature. The methodology and reporting of the review were guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines. The methodological framework for this scoping review was developed before the initiation of the literature search and made available on the Open Science Framework to enhance transparency. The procedures defined in the protocol were followed throughout the review process.

### Identifying the Research Questions

The following research questions guided the review:

- RQ1. In which assessment domains of pediatric occupational therapy have AI-based approaches been applied?
- RQ2. What types of AI methods have been used across these assessment domains in pediatric occupational therapy?
- RQ3. How have the outcomes and contributions of AI-based assessment applications been reported in the existing literature?
- RQ4. Which pediatric populations have been represented in studies using AI-based assessment approaches in pediatric occupational therapy?

### Search Strategy

A comprehensive literature search was conducted to identify studies examining the use of AI-based approaches for assessment purposes in pediatric occupational therapy. The following electronic databases were searched: CINAHL (through EBSCOhost), OTseeker, PubMed/MEDLINE, Scopus, and Web of Science.

The search strategy combined relevant keywords and, where applicable, controlled vocabulary terms related to AI, assessment, pediatric populations, and occupational therapy. Search terms included combinations of "artificial intelligence," "machine learning," "deep learning," "computer vision," "algorithm," "pattern recognition," "intelligent systems," "assessment," "evaluation," "measurement," "screening," "functional assessment," "functional analysis," "performance-based assessment," "occupational therapy," "pediatric," "child," "infant," and "adolescent." Boolean operators (AND/OR) were

used to combine search terms appropriately. The search strategy was adapted for each database to account for differences in indexing and search functionalities.

No restrictions were applied regarding geographical location. The search was filtered to English peer-reviewed articles published between 2015 and 2025. In addition to electronic database searching, the reference lists of included studies were manually screened to identify additional relevant publications.

The search strategy was reviewed and refined by the research team before implementation to ensure adequate coverage and relevance. All records retrieved from the database searches were imported into EndNote, a reference management software, which was used to organize citations and identify and remove duplicate records before the screening process.

### Study Selection

Study selection was conducted in two stages, beginning with title and abstract screening, followed by full-text review. Three researchers independently screened the titles and abstracts of all retrieved records to assess relevance based on the predefined inclusion and exclusion criteria. Studies were considered eligible if they reported empirical research examining AI-based approaches used for assessment purposes in pediatric occupational therapy. Eligible studies included qualitative, quantitative, and mixed-methods research, published in English in peer-reviewed journals between 2015 and 2025. Studies were excluded if they focused on AI in health care without direct relevance to occupational therapy; did not address assessment processes; were systematic reviews, scoping reviews, or narrative reviews; did not involve human participants; examined assistive technologies without an AI component; or consisted of opinion pieces, editorials, or conceptual papers lacking empirical data. Full-text versions of studies deemed potentially eligible were retrieved and independently assessed by the same researchers for final inclusion. The reference lists of all included articles were also screened for additional eligible studies not captured in the primary database search. Any discrepancies in screening decisions were resolved through discussion until consensus was achieved. The study selection process is summarized in a PRISMA-ScR flow diagram (Fig. 1), detailing the number of records identified, screened, included, and excluded at each stage.

### Data Charting

A standardized data charting form was developed to systematically extract relevant information from the

included studies. The data charting framework was designed to align with the objectives of the scoping review and the predefined research questions. Before full data charting, the form was reviewed by the research team to ensure clarity and consistency. For each included study, the following information was charted: Publication characteristics (author(s), year of publication, and country), study design, participant characteristics (age range and pediatric population when reported), type of AI method used, assessment domain(s) addressed within pediatric occupational therapy, purpose of the AI-based assessment application, and how assessment-related outcomes were reported in the literature.

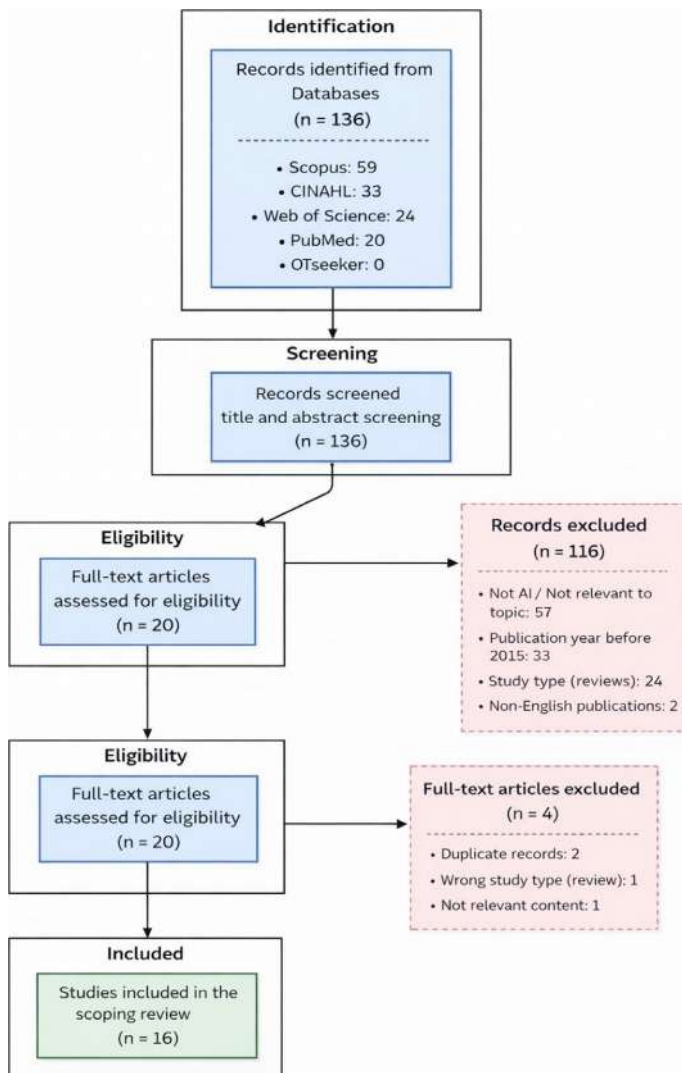
Assessment domains were defined a priori to support systematic mapping and included sensory-perceptual, motor, cognitive, psychosocial, environmental, ADL, and participation-related domains, consistent with occupational therapy assessment frameworks.<sup>[10]</sup> Studies could be assigned to more than one assessment domain where applicable.

Data charting was performed independently by three researchers. Any discrepancies in the charted data were discussed and resolved through consensus to ensure accuracy and consistency. The extracted data were subsequently used to support descriptive synthesis and mapping of assessment domains, AI methods, and pediatric populations addressed in the included studies.

### Data Synthesis and Analysis

Data synthesis was performed using a descriptive and narrative approach, in accordance with the objectives and methodological guidance of the PRISMA-ScR. The purpose of the synthesis was to map the breadth, characteristics, and distribution of existing evidence, rather than to evaluate effectiveness or conduct quantitative pooling of results. Following data charting, the extracted data were organized and summarized to address the predefined research questions. Studies were grouped according to assessment domains in pediatric occupational therapy, types of AI approaches, data modalities, reported outcome measures, and pediatric populations represented. Assessment domains were categorized a priori (sensory-perceptual, motor, cognitive, psychosocial, environmental, ADL, and participation), and studies addressing multiple domains were classified under all relevant categories.

AI approaches were descriptively classified (e.g., machine learning, deep learning, computer vision, hybrid methods), and their distribution across assessment domains was



**Figure 1.** Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews flow diagram illustrating the study selection process.

examined. Data synthesis was supported by tabular and visual representations, including summary tables and a heatmap illustrating the relationship between AI techniques and assessment domains. The findings were synthesized narratively, with an emphasis on overall patterns, trends, and gaps in the literature, consistent with the exploratory aims of a scoping review.

## Results

An initial search identified 136 articles. After removing duplicates and conducting detailed screening, 20 studies were deemed potentially relevant. Applying the inclusion criteria narrowed this down to 16 articles included in the qualitative synthesis. The screening process is outlined in Figure 1, following PRISMA-ScR guidelines.

The characteristics of the included studies are summarized in Table 1, providing an overview of study designs, AI techniques, data modalities, assessment domains, outcome measures, and population characteristics.

## Study Design

Most included studies employed observational designs and primarily focused on validation and measurement purposes (Table 1). Cross-sectional approaches were commonly reported, whereas longitudinal and experimental designs were less frequently observed.

## AI Type

Machine learning emerged as the predominant AI approach across the included studies (Table 1). Deep learning and computer vision techniques were also frequently applied, while hybrid methods and rule-based systems were used only in a limited number of studies. No studies reported the use of natural language processing techniques.

## Data Modality

A wide range of data modalities was used across studies, with image- and video-based data being most common (Table 1). Clinical records, questionnaire-based data, and wearable sensors were also utilized, whereas depth cameras, audio data, and multimodal data integration were relatively uncommon.

## Outcome Measures

Accuracy emerged as the most frequently reported outcome measure across the included studies, followed by validity or correlation metrics and clinician agreement (Table 1). Sensitivity, specificity, and area under the curve were also commonly used performance indicators, whereas F1-score and precision/recall metrics were reported less frequently. Reliability indices, such as intraclass correlation coefficients, were included in a limited number of studies, and measures related to time efficiency were reported only occasionally. Other advanced performance indicators, including mean absolute error, root mean squared error, and  $R^2$ , were rarely reported.

## Population Characteristics

The included studies predominantly involved general pediatric populations and children with neurological or developmental conditions, particularly cerebral palsy (Table 1). A smaller number of studies focused on children with developmental coordination disorder, brain tumor survivors, and specific learning disorders, while a limited number included mixed clinical populations or children with idiopathic toe walking.

**Table 1.** Characteristics of included studies (n=16)

	n	%
Study design		
Observational	15	93.8
Experimental (RCT)	1	6.3
Validation/measurement	13	81.3
Feasibility/pilot	1	6.3
Cross-sectional	11	68.8
Longitudinal	4	25.0
AI type		
Machine learning	14	87.5
Deep learning	6	37.5
Computer vision	6	37.5
Hybrid	3	18.8
Rule-based/expert system	1	6.3
Data modality		
Video	5	31.3
Image/photo	7	43.8
Wearable sensors (IMU)	4	25.0
Depth camera (Kinect)	1	6.3
Audio	1	6.3
EHR/clinical records	5	31.3
Questionnaire/scale	5	31.3
Multi-modal	2	12.5
Other	1	6.3
Assessment domain		
Motor	14	87.5
Sensory-perceptual	9	56.2
Cognitive	2	12.5
ADL	2	12.5
Outcome type		
Accuracy	13	81.3
Sensitivity	6	37.5
Specificity	6	37.5
AUC	6	37.5
F1-score	3	18.8
Precision/recall	2	12.5
Validity/correlation	10	62.5
Reliability (ICC)	3	18.8
Agreement (clinician)	8	50.0
Time/efficiency	4	25.0
Other (MAE, RMSE, R <sup>2</sup> )	1	6.3
Population group		
Cerebral palsy	5	31.25
Developmental coordination disorder	1	6.3
Brain tumor survivors	1	6.3
General pediatric	6	37.5
Specific learning disorders	1	6.3
Other (mixed, ITW, etc.)	2	12.5

AUC: Area under the curve; EHR: Electronic health records; IMU: Inertial measurement unit; RCT: Randomized controlled trials; AI: Artificial intelligence; ADL: Activities of daily living; MAE: Mean absolute error; RMSE: Root mean squared error; ICC: Intraclass correlation coefficient. Categories are not mutually exclusive; therefore, percentages may exceed 100%.

## Assessment Domains and AI Techniques

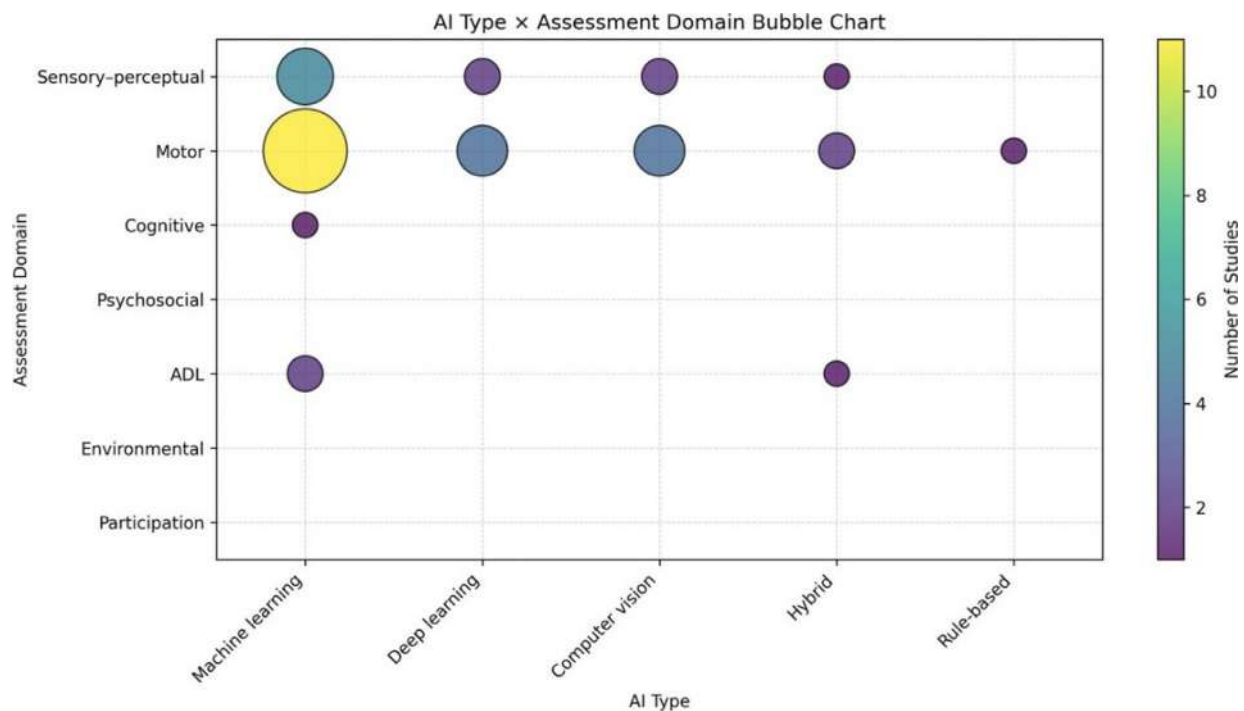
Figure 2 illustrates the distribution of AI types across assessment domains. Most studies primarily focused on motor functioning (n=14/16, 87.5%), followed by sensory-perceptual processes (n=9/16, 56.2%). Machine learning was predominantly applied in motor-related assessments and, to a lesser extent, in sensory-perceptual evaluations. Deep learning and computer vision methods were also mainly used in these two domains. Only a limited number of studies addressed cognitive functioning and ADL (each 12.5%). Notably, none of the included studies evaluated psychosocial, environmental, or participation-related domains. Several studies employed multiple AI techniques and addressed more than one assessment domain. Therefore, these studies were classified under all relevant categories in the heatmap analysis.

Detailed characteristics of individual studies, including AI methods, assessment tools, and outcome measures, are provided in Appendix 1.

## Discussion

The findings of this scoping review provide a comprehensive synthesis of how AI-based assessment approaches have been applied within pediatric occupational therapy. The existing literature demonstrates a clear concentration of AI applications within motor and sensory-perceptual assessment domains, while cognitive and ADL-related assessments remain comparatively underrepresented. Notably, no studies were identified that explicitly addressed psychosocial, environmental, or participation-related domains. From an occupational therapy perspective, these findings highlight a critical imbalance between performance-based and occupation-centered assessment approaches. According to the occupational therapy practice framework-4, assessment should encompass not only performance skills (e.g., motor and process skills) but also participation, contextual factors, and environmental influences. However, the current distribution of AI-based applications appears to be heavily skewed toward observable motor performance components, with limited attention to broader occupational engagement and real-life participation contexts.

In line with this perspective, two key findings emerged from this review. First, AI-based assessment applications in pediatric occupational therapy are predominantly concentrated in motor and sensory-perceptual domains, which are more easily quantifiable through image- and sensor-based data. Second, there is a notable absence of



**Figure 2.** Distribution of artificial intelligence types across assessment domains.

AI applications targeting psychosocial, environmental, and participation-related domains, despite their central importance in occupational therapy practice. These findings are consistent with existing literature on AI applications in rehabilitation, which similarly emphasizes the predominance of performance-based approaches and the early-stage integration of AI into clinical practice.<sup>[11,12]</sup> Previous reviews have also highlighted limitations in methodological rigor and validation processes, including the limited use of multimodal data and external validation strategies.<sup>[13]</sup> In this context, the present study extends the literature by specifically mapping these patterns within pediatric occupational therapy and identifying critical gaps in participation and environmental domains.

The predominance of observational and validation-focused study designs further indicates that AI technologies in this field are primarily positioned as measurement-enhancement tools, rather than as fully integrated components of routine clinical assessment and decision-making. This methodological pattern aligns with trends reported in broader pediatric health and rehabilitation research, where exploratory and validation-oriented designs are commonly adopted during the early phases of technological implementation.<sup>[14,15]</sup> While such approaches are valuable for establishing feasibility, proof of concept, and clinical relevance, they inherently limit causal inference and the generalizability of findings.

The limited representation of experimental randomized controlled trials (RCTs) identified in this review highlights a critical methodological gap. As RCTs are widely regarded as providing higher levels of evidence through controlled and structured designs, their underrepresentation raises concerns regarding potential bias and the robustness of current findings.<sup>[16]</sup> Consistent with previous critiques in the literature, there is a clear need for more rigorously designed experimental studies to systematically evaluate the efficacy, safety, and clinical impact of AI-based assessment approaches in pediatric occupational therapy.<sup>[15,17]</sup> Addressing this gap may strengthen evidence-based integration and more effectively inform clinical decision-making processes.

The dominance of machine learning techniques, including deep learning and computer vision-based methods, observed in this review is consistent with broader trends in AI-driven health research. These approaches are particularly well-suited for tasks such as image classification, pattern recognition, and movement analysis, which are central to motor and sensory-perceptual assessments commonly employed in pediatric populations.<sup>[14,18–20]</sup> The use of hybrid models further reflects an emerging tendency to integrate multiple AI techniques to improve model performance, adaptability, and applicability across diverse assessment contexts.<sup>[21,22]</sup>

In parallel, the strong reliance on image- and photo-based data underscores a broader shift toward visually driven assessment approaches in pediatric rehabilitation settings.

This trend is supported by evidence demonstrating that computer vision-based techniques can enhance objectivity and precision in the evaluation of motor performance and developmental outcomes in children.<sup>[14,15,23]</sup> Nevertheless, the comparatively limited use of wearable sensors and the infrequent implementation of multimodal data integration highlight significant opportunities for methodological advancement. Emerging research suggests that combining visual data with physiological and kinematic information may yield more comprehensive and ecologically valid assessments, thereby improving predictive accuracy and clinical relevance.<sup>[24–27]</sup>

While the emphasis on motor functioning reflects clinical priorities, it also reveals a notable imbalance in the literature.<sup>[15,28]</sup> Specifically, limited attention has been given to cognitive functioning and ADL, alongside a conspicuous absence of AI-based assessment approaches targeting participation and environmental domains. This gap is particularly salient given that participation and person-environment interactions constitute foundational constructs within pediatric occupational therapy practice. This imbalance may be explained by the inherent complexity of participation and psychosocial constructs, which are context-dependent, subjective, and difficult to operationalize using current AI methodologies. In contrast, motor performance can be more readily captured through structured, quantifiable data such as movement kinematics or visual recordings. This technological bias may inadvertently reinforce a reductionist perspective, limiting the ability of AI systems to reflect the holistic and occupation-centered nature of pediatric occupational therapy. These applications remain largely confined to performance-level indicators, with broader, contextually embedded aspects of occupational engagement remaining underrepresented. Expanding AI-based assessment approaches to encompass participation and environmental factors may therefore support more holistic, occupation-centered, and family-centered evaluations.

<sup>[29,30]</sup> From a clinical perspective, the current emphasis on performance-level indicators raises concerns regarding the alignment of AI-based assessments with the core principles of occupational therapy, which prioritize participation, context, and meaningful engagement in daily life. This limitation may reduce the ecological validity of assessments and their relevance for clinical decision-making. Furthermore, although accuracy was the most frequently reported outcome measure, the limited use of reliability indices and advanced performance metrics constrains the clinical interpretability of existing tools. Broader outcome reporting, combined with assessments

that capture participation and environmental context, may enhance the robustness and ecological validity of AI-based assessment systems.<sup>[22,24,25]</sup> In practical terms, current AI-based approaches may support clinicians in tasks such as objective motor performance analysis, early screening, and progress monitoring. For example, computer vision-based systems can assist in quantifying movement patterns, while wearable sensor-based approaches may provide continuous data on functional performance. However, expanding these applications toward real-world contexts and participation-based outcomes is essential to better support occupation-centered clinical decision-making.

Another important issue identified in this review relates to the limited transparency and variability in reporting AI model development processes across the included studies. In several cases, insufficient detail was provided regarding model architecture, training procedures, parameter selection, and validation strategies, which may hinder reproducibility and critical appraisal. Furthermore, although performance metrics such as accuracy were frequently reported, their clinical relevance and interpretability in occupational therapy contexts were often unclear.

Finally, the predominant focus on general pediatric populations and conditions such as cerebral palsy mirrors prevailing priorities in pediatric rehabilitation research, where these groups are frequently targeted due to their prevalence and clinical significance. However, expanding future research to include more diverse and mixed clinical populations may enhance the generalizability and applicability of AI-based assessment approaches across a wider range of pediatric profiles.<sup>[23,31]</sup> Such diversification holds the potential to support more personalized assessment strategies and contribute to improved outcomes for currently underrepresented pediatric populations.<sup>[14,29]</sup>

In addition to these considerations, ethical aspects should also be acknowledged in the use of AI in pediatric populations. Ethical considerations are important in the use of AI in pediatric populations, particularly regarding data privacy, informed consent, and the protection of sensitive child data. Transparency and explainability are also essential to ensure trust in clinical practice. AI should be considered as a decision-support tool that complements, rather than replaces, clinical judgment.

This review has several strengths. It provides a comprehensive and systematic mapping of AI-based assessment approaches across multiple domains in pediatric occupational therapy. In addition, the use of a structured framework to categorize assessment domains allows for a clearer identification of trends and gaps in the literature.

## Study Limitations

Despite these contributions, several limitations should be considered. Although a comprehensive and systematically developed search strategy was implemented across multiple major databases, relevant studies may not have been identified due to variations in terminology, indexing practices, or publication in sources not covered by the selected databases. The restriction to English-language, peer-reviewed publications may have resulted in the exclusion of relevant studies published in other languages or disseminated through gray literature, potentially limiting the global representation of AI-based assessment practices. In addition, the review was confined to studies that explicitly reported AI-based approaches within pediatric occupational therapy contexts; therefore, relevant assessment applications described in adjacent disciplines but not clearly framed within occupational therapy may not have been captured. In addition, variability and limited transparency in reporting AI methodologies across the included studies may have constrained the depth of analysis.

## Conclusion

This scoping review reveals a clear imbalance in the current use of AI-based assessment approaches within pediatric occupational therapy. Existing applications are largely concentrated in performance-oriented domains, particularly motor and sensory-perceptual functions, while key occupational therapy domains such as participation, psychosocial functioning, and environmental context remain largely unexplored.

This imbalance highlights a critical gap between the capabilities of current AI technologies and the holistic, occupation-centered perspective that underpins pediatric occupational therapy practice. As a result, the clinical relevance and ecological validity of existing AI-based assessment tools may be limited.

Future research should focus on developing AI-based assessment approaches that extend beyond performance-level indicators to incorporate participation, contextual, and environmental factors. Integrating multimodal data sources and adopting more rigorous and transparent methodological practices – including detailed reporting of training, validation (e.g., cross-validation or external validation), and testing procedures – may enhance both the clinical relevance and reliability of these systems. Advancing AI applications toward more comprehensive and contextually grounded frameworks may better support clinical decision-making and promote more client-centered pediatric occupational therapy practice.

**Ethics Committee Approval:** Ethical approval was not required for this study since this is a review article.

**Conflict of Interest:** The authors declare that there is no conflict of interest.

**Financial Disclosure:** The authors declared that this study received no financial support.

**Use of AI for Writing Assistance:** The authors did not use AI-assisted writing tools.

**Authorship Contributions:** Concept: ZK; Design: ZK; Supervision: ZK; Data Collection and/or Processing: BEA; Analysis and/or Interpretation: DMA; Literature Search: ZK, BEA, DMA; Writing: ZK, BEA, DMA; Critical Reviews: ZK.

**Peer-review:** Double blind peer-reviewed.

## References

1. Beets B, Newman TP, Howell EL, Bao L, Yang S. Surveying public perceptions of artificial intelligence in health care in the united states: systematic review. *J Med Internet Res* 2023;25:e40337. [\[CrossRef\]](#)
2. Matmi MM, Shahba S, Alrwuili AA, Alotaibi MM, Alayli MH, Asiri AM, et al. Application of artificial intelligence in community-based primary health care: systematic review. *Journal of Namibian Studies* 2023;35:1269-92.
3. Bhatt P, Liu J, Gong Y, Wang J, Guo Y. Emerging artificial intelligence-empowered mhealth: scoping review. *JMIR Mhealth Uhealth* 2022;10(6):e35053. [\[CrossRef\]](#)
4. Mousavi Baigi SF, Dahmardeh Kemmak F, Sarbaz M, Norouzi Aval R, Kimiafar Kh. Application of Artificial Intelligence in Occupational Therapy. *Health Education and Health Promotion* 2024;12(3):513-20.
5. Sharma N. Use of AI in pediatric occupational therapy: A review. *Int J Cur Res Rev* 2024;16(2):1-6 [\[CrossRef\]](#)
6. Skuthan A, Stav W. Profile of assessments in pediatric occupational therapy practice. *Am J Occup Ther* 2023;77(5):7705205010. [\[CrossRef\]](#)
7. Hagihara H, Ienaga N, Enomoto D, Takahata S, Ishihara H, Noda H, et al. Computer vision-based approach for quantifying occupational therapists' qualitative evaluations of postural control. *Occup Ther In* 2020;2020(1):8542191. [\[CrossRef\]](#)
8. Tsai YT, Lee JS, Huang CY. Research on Applying Deep Learning to Visual-Motor Integration Assessment Systems in Pediatric Rehabilitation Medicine. *Algorithms*, 2024;17(9):413. [\[CrossRef\]](#)
9. Kokkotis C, Kansizoglou I, Stampoulis T, Giannakou E, Siaperas P, Kallidis S, et al. Artificial Intelligence as Assessment Tool in Occupational Therapy: A Scoping Review. *BioMedInformatics* 2025;5(2):22. [\[CrossRef\]](#)
10. American Occupational Therapy Association. AOTA 2021 Standards for Continuing Competence in Occupational Therapy. *Am J Occup Ther* 2022;75(Supp. 3):7513410040. [\[CrossRef\]](#)
11. Lanotte F, O'Brien MK, Jayaraman A. AI in Rehabilitation medicine: opportunities and challenges. *Ann Rehabil Med* 2023;47(6):444-58. [\[CrossRef\]](#)

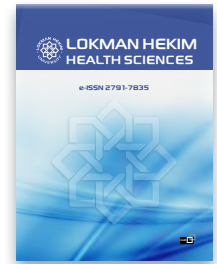
12. Alshami A, Nashwan A, AlDardour A, Qusini A. Artificial Intelligence in rehabilitation: A narrative review on advancing patient care. *Rehabilitacion (Madr)* 2025;59(2):100911. [\[CrossRef\]](#)
13. Morone G, Carbonetti R, Martino Cinnera A, Malfitano C, Bisirri A, Negrini F, et al. Artificial intelligence in rehabilitation: a living systematic mapping review - first release. *Eur J Phys Rehabil Med* 2025;61(5):694-708. [\[CrossRef\]](#)
14. Gunturkun F, Bakir-Batu B, Siddiqui A, Lakin K, Hoehn ME, Vestal R, et al. Development of a deep learning model for retinal hemorrhage detection on head computed tomography in young children. *JAMA Netw Open* 2023;6(6):e2319420. [\[CrossRef\]](#)
15. Shaari H, Kevrić J, Jukić S, Bešić L, Jokić D, Ahmed N, et al. Deep learning-based studies on pediatric brain tumors imaging: narrative review of techniques and challenges. *Brain Sci* 2021;11(6):716. [\[CrossRef\]](#)
16. Aguilar-Ortega R, Berral-Soler R, Jiménez-Velasco I, Romero-Ramírez FJ, García-Marín M, Zafra-Palma J et al., UCO Physical rehabilitation: new dataset and study of human pose estimation methods on physical rehabilitation exercises. *Sensors* 2023;23(21):8862. [\[CrossRef\]](#)
17. Moro M, Marchesi G, Hesse F, Odone F, Casadio M. Markerless vs. marker-based gait analysis: a proof of concept study. *Sensors (Basel)* 2022;22(5):2011. [\[CrossRef\]](#)
18. Liu Y, Li C, Li F, Lin R, Zhang D, Lian Y. Advances in computer vision and deep learning-facilitated early detection of melanoma. *Briefings in Functional Genomics* 2025;24:elaf002. [\[CrossRef\]](#)
19. Xiong J, Lu L, Wang H, Yang J, Gui AG. Object-level trajectories based fine-grained action recognition in visual iot applications. *IEEE Access* 2019;7:103629-38. [\[CrossRef\]](#)
20. Rames JD, Tunaboynu MF, Emanuels AF, Moran SL. From theory to practice: moving toward artificial intelligence-powered computer vision applications for peripheral motor nerve assessment of the hand. *Plast Reconstr Surg Glob Open* 2025;13(4):e6674. [\[CrossRef\]](#)
21. Zhou J, Luo Y, Darcy JW, Lafata KJ, Ruiz JR, Grego S. Long-term, automated stool monitoring using a novel smart toilet: A feasibility study. *Neurogastroenterol Motil* 2025;37(1):e14954. [\[CrossRef\]](#)
22. Evain E, Faraz K, Grenier T, Garcia D, De Craene M, Bernard O. A pilot study on convolutional neural networks for motion estimation from ultrasound images. *IEEE Trans Ultrason Ferroelectr Freq Control* 2020;67(12):2565-73. [\[CrossRef\]](#)
23. Gao Y, Jiang Y, Peng Y, Yuan F, Zhang X, Wang J. Medical image segmentation: a comprehensive review of deep learning-based methods. *Tomography* 2025;11(5):52. [\[CrossRef\]](#)
24. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med* 2022;28(9):1773-84. [\[CrossRef\]](#)
25. Zheng C, Wu W, Chen C, Yang T, Zhu S, Shen J, et al. Deep learning-based human pose estimation: a survey. 2020;56(1):1-37. [\[CrossRef\]](#)
26. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med* 2016;375(13):1216-9. [\[CrossRef\]](#)
27. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44-56. [\[CrossRef\]](#)
28. Ray LSS, Zhou B, Suh S, Lukowicz P. A comprehensive evaluation of marker-based, markerless methods for loose garment scenarios in varying camera configurations. *Frontiers in Computer Science* 2024;6:1379925 [\[CrossRef\]](#)
29. Anwar A, Kanwal S, Tahir M, Saqib M, Uzair M, Imam Rahmani MK, et al. Image aesthetic assessment: a comparative study of hand-crafted & deep learning models. *IEEE Access* 2022;10:101770-89. [\[CrossRef\]](#)
30. Usman M, Zhong J. Skeleton-based motion prediction: A survey. *Front Comput Neurosci* 2022;16:1051222. [\[CrossRef\]](#)
31. Chong PL, Vaigeshwari V, Mohammed Reyasudin BK, Noor Hidayah BRA, Tatchanaamoorti P, Yeow JA, et al. Integrating artificial intelligence in healthcare: applications, challenges, and future directions. *Future Sci OA* 2025;11(1):2527505. [\[CrossRef\]](#)
32. Fair-Field T, Modayur B. Clinical validation of an abridged AIMS: Streamlining motor screening in the first-year infant. *Early Hum Dev* 2025;202:106207. [\[CrossRef\]](#)
33. Lin P-H, Kuo P-H. Ensemble learning based functional independence ability estimator for pediatric brain tumor survivors. *Health Informatics Journal*. 2022;28(4). [\[CrossRef\]](#)
34. Philip BA, Li F, Hawkins-Chernof E, Chen L, Swamidass V, Zwir I. Motor assessment with the STEGA iPad app to measure handwriting in children. *Am J Occup Ther* 2023;77(3):7703205010. [\[CrossRef\]](#)
35. Khaksar S, Pan H, Borazjani B, Murray I, Agrawal H, Liu W, et al. Application of inertial measurement units and machine learning classification in cerebral palsy: randomized controlled trial. *JMIR Rehabil Assist Technol* 2021;8(4):e29769. [\[CrossRef\]](#)
36. Li R, Fu H, Zheng Y, Lo WL, Yu JJ, Sit CHP, et al. Automated fine motor evaluation for developmental coordination disorder. *IEEE Trans Neural Syst Rehabil Eng* 2019;27(5):963-73. [\[CrossRef\]](#)
37. Duran I, Stark C, Saglam A, Semmelweis A, Lioba Wunram H, Spiess K, et al. Artificial intelligence to improve efficiency of administration of gross motor function assessment in children with cerebral palsy. *Dev Med Child Neurol* 2022;64(2):228-34. [\[CrossRef\]](#)
38. Syafrudin M, Alfian G, Fitriyani NL, Anshari M, Hadibarata T, Fatwanto A, et al. A self-care prediction model for children with disability based on genetic algorithm and extreme gradient boosting. *Mathematics* 2020;8(9):1590. [\[CrossRef\]](#)
39. Soangra R, Wen Y, Yann H, Grant-Beuttler M. Classifying toe walking gait patterns among children diagnosed with idiopathic toe walking using wearable sensors and machine learning algorithms. *IEEE Access* 2022;10:77054-67. [\[CrossRef\]](#)
40. Yiwen W, Yonghui Y. Development and validation of a prognostic model for independent walking in children with cerebral palsy based on machine learning. *Arch Phys Med Rehabil* 2025;106(12):1850-8. [\[CrossRef\]](#)
41. Hegde N, Zhang T, Uswatte G, Taub E, Barman J, McKay S, et al. The pediatric smartshoe: wearable sensor system for ambulatory monitoring of physical activity and gait. *IEEE Trans Neural Syst Rehabil Eng* 2018;26(2):477-86. [\[CrossRef\]](#)

42. Rico-Olarte C, LÓPEZ DG, Becker L, Eskofier B. Towards classifying cognitive performance by sensing electrodermal activity in children with specific learning disorders. *IEEE Access* 2020;8:196187-96. [\[CrossRef\]](#)
43. Schafmeyer L, Losch H, Bossier C, Lanzl W, Wunram HL, Schoenau E, et al. Using artificial intelligence-based technologies to detect clinically relevant changes of gross motor function in children with cerebral palsy. *Dev Med Child Neurol* 2024;66(2):226-32. [\[CrossRef\]](#)
44. Ienaga N, Takahata S, Terayama K, Enomoto D, Ishihara H, Noda H, et al. Development and verification of postural control assessment using deep-learning-based pose estimators: towards clinical applications. *Occup Ther Int* 2022;2022:6952999. [\[CrossRef\]](#)
45. Villegas-Ch W, Urbina-Camacho I, García-Ortiz J. Detection of abnormal patterns in children's handwriting by using an artificial-intelligence-based method. *Informatics* 2023;10:52. [\[CrossRef\]](#)

## Appendix 1. Detailed characteristics of included studies

Author(s), year	Title	Journal	Country/region	Study design	Setting/context	Age range	Pediatric population/condition	Sample size (n)	Study objective/aim	AI method type	AI model/algorithm	Data modality	Assessment domain(s)	Assessment tool/measure	Purpose in assessment	Clinical application/intended Use	Reported outcomes/metrics	Key findings (assessment)	Reported limitations
Fair-Field and Modayur <sup>[32]</sup>	Clinical validation of an abridged AIMS: Streamlining motor screening in the 1st-year infant	Early Human Development	USA	Observational, validation	Clinical and telehealth video-based screening	0–12 months	Typically developing and at-risk infants	123 (102 training, 21 validation)	To validate a shortened version of the AIMS using ML for early motor screening	ML	SVR	Video	Motor	AIMS	Early identification of motor delay	Screening and early detection	Accuracy, sensitivity, specificity, ICC, correlation, time efficiency	The abridged 15-item AIMS showed excellent agreement with the full scale (r=0.99), high sensitivity (1.00), and reduced assessment time	Small validation sample, limited generalizability, reliance on video recordings
Lin and Kuo <sup>[33]</sup>	Ensemble learning-based functional independence ability estimator for pediatric brain tumor survivors	Health Informatics Journal	Taiwan	Observational, Validation	Rehabilitation clinic/ Secondary dataset	6–18 years	Pediatric brain tumor survivors	73	To predict functional independence (WeeFIM) scores using ensemble ML models	ML (Ensemble)	RF, MLP, AdaBoost, SVR, DT, k-NNR	EHR/Clinical records, Questionnaire/Scale	ADL, Motor	WeeFIM, WISC-IV	Estimation and prediction of functional independence scores	Prognosis and rehabilitation planning	MAE, RMSE, R2	Ensemble models demonstrated superior predictive performance compared with single ML models and accurately estimated WeeFIM total, cognition, mobility, and self-care scores	Secondary dataset use, limited sample size, lack of external validation, and limited generalizability
Philip et al. <sup>[34]</sup>	Motor Assessment with the STEGA iPad App to Measure Handwriting in Children	American Journal of Occupational Therapy	USA	Observational, Validation	School and clinical settings	6–11 years	Typically developing children and children with handwriting difficulties	57	To evaluate the validity of the STEGA iPad app for handwriting assessment	ML	SVR	Image/Tablet input	Motor, Sensory-perceptual	STEGA iPad App, DASH	Objective measurement of handwriting performance	Screening and progress monitoring	Accuracy, validity coefficients, and agreement	The STEGA app demonstrated strong validity and reliability for assessing handwriting speed, pressure, and legibility	Limited sample size, restricted age range, use of tablet-based writing only
Khaksar et al. <sup>[35]</sup>	Application of inertial measurement units and ML classification in cerebral palsy: Randomized controlled trial	JMIR Rehabilitation and Assistive Technologies	Australia	RCT, observational	Multicenter clinical rehabilitation settings	5–15 years	Children with and without cerebral palsy	190	To develop and evaluate an IMU-based system using ML to classify wrist movement patterns in children with cerebral palsy	ML	RF, C4.5, SVM, KNN, MLP, Naïve Bayes, Logistic Regression	Wearable sensors (IMU - Accelerometer and Gyroscope)	Motor, Sensory-perceptual	Custom IMU system, Goniometer, Vicon	Objective measurement of active wrist range of motion and movement classification	Therapy evaluation and long-term monitoring	Accuracy, AUC, ROC, classification rate	RF and C4.5 achieved high classification accuracy (up to 89.39%) in distinguishing cerebral palsy and typical movement patterns	Sensor drift, data noise, technical issues, limited comfort, potential data loss, and restricted ecological validity
Li et al. <sup>[36]</sup>	Automated fine motor evaluation for developmental coordination disorder	IEEE Transactions on Neural Systems and Rehabilitation Engineering	Hong Kong/China	Observational, validation	Laboratory and clinical assessment setting	6–10 years	Children with DCD and TD	40 (20 DCD, 20 TD)	To develop an automated system for assessing fine motor performance in children with DCD	ML	CNN-based temporal filtering, image processing	Motion capture/digital pen, depth camera (e.g., Kinect)	Motor, Sensory-perceptual	Based on MABC-2 tasks (drawing trail, Threading lace, Posting coins)	Automated scoring to replace subjective observational analysis	Screening and diagnostic support	Accuracy, sensitivity, specificity, classification rate	The automated system successfully differentiated children with DCD from controls with high classification accuracy	Small sample size, controlled laboratory setting, limited ecological validity
Duran et al. <sup>[37]</sup>	Artificial intelligence to improve the efficiency of the administration of gross motor function assessment in children with cerebral palsy	Developmental Medicine and Child Neurology	Germany, Turkey	Retrospective, observational, validation	Single-center rehabilitation program ("Auf die Beine")	<18 years (mean age 8 years 10 months)	Children with cerebral palsy	1217 unique assessments (validation set: 187 assessments)	To create a reduced version of the GMFM-66 (rGMFM-66) to improve clinical efficiency	ML	RF, SVM, feed-forward neural network	Clinical assessment records, Questionnaire/Scale	Motor	GMFM-66, rGMFM-66, GMFCS	Predicting the total score from a subset of items to save time	Outcome monitoring and therapy evaluation	ICC, MAE, Bland-Altman, sensitivity, specificity	The rGMFM-66 showed excellent agreement with the full GMFM-66 (ICC>0.99) and significantly reduced assessment time	Retrospective design, selection bias, limited representation of GMFCS I and V, lack of prospective validation
Hagihara et al. <sup>[7]</sup>	CV-based approach for quantifying occupational therapists' qualitative evaluations of postural control	Occupational Therapy International	Japan	Observational, validation	Nursery schools	3–6 years	Typically developing preschool children	34	To develop and validate a CV-based method for quantifying postural control based on therapists' qualitative evaluations	Deep learning/CV	Pose estimation-based CV (OpenPose)	Video	Motor, Sensory-perceptual	JPAN (One Arm and One Leg Balance task), TQCE	Quantification of postural control performance	Screening and detailed motor assessment	Spearman correlation, regression coefficients, ICC	CV-based indices (SPB, AG) showed significant correlations with therapists' evaluations and outperformed conventional duration-based measures	Small sample size, typically developing sample only, limited clinical diversity, single-task assessment
Syafrudin et al. <sup>[38]</sup>	A self-care prediction model for children with disabilities based on a genetic algorithm and extreme gradient boosting	Mathematics	South Korea, Brunei, Malaysia, Indonesia	Observational, validation	Secondary dataset analysis (SCADI, ICF-CY)	6–18 years	Children with various disabilities	70	To develop a hybrid prediction model (GA-XGBoost) for classifying self-care activity levels	ML (hybrid feature selection)	Genetic algorithm+XGBoost	EHR/questionnaire-based dataset	ADL	ICF-CY	Automated classification of self-care difficulties	Diagnostic support and rehabilitation planning	Accuracy, precision, recall, F1-score, AUC	GA-XGBoost outperformed other ML models, achieving up to 98.57% accuracy in binary classification and 90% in multi-class tasks	Use of secondary dataset, small sample size, lack of real-time clinical validation, and limited ecological validity
Soangra et al. <sup>[39]</sup>	Classifying Toe Walking Gait Patterns Among Children Diagnosed With Idiopathic Toe Walking Using Wearable Sensors and Machine Learning Algorithms	IEEE Access	USA	Observational, Validation	Laboratory and outpatient gait analysis setting	4–12 years	Children with idiopathic toe walking (ITW)	35 (17 ITW, 18 Typically Developing)	To identify and classify toe-walking steps from typical gait using wearable inertial sensors and ML.	Machine Learning	Random Forest, SVM, KNN, Decision Tree	Wearable sensors (IMU)	Motor, Sensory-perceptual	Custom IMU system, gait analysis protocol	Automated detection of abnormal gait cycles (toe-walking steps)	Clinical decision support, monitoring treatment (e.g., botox, casting) and intervention planning.	Accuracy, precision, recall, F1-score	ML models achieved high accuracy (>90%) in distinguishing ITW from typical gait patterns	Small sample size, controlled environment, limited generalizability, short recording duration
Yiwen & Yonghui <sup>[40]</sup>	Development and Validation of a Prognostic Model for Independent Walking in Children With Cerebral Palsy Based on Machine Learning	Archives of Physical Medicine and Rehabilitation	China	Retrospective cohort, Observational, Validation	National CP registry and telephone follow-up	Birth–6 years (followed up to 6 years)	Children with cerebral palsy	807	To develop and validate prognostic models for predicting independent walking ability before age 6.	Machine Learning	Logistic Regression, XGBoost, MLP, RF, GBM, SVM, KNN, Naïve Bayes	EHR / Registry data, Questionnaire	Motor	GMFCS, GMFM-88, structured interview	Prediction of future walking ability	Prognosis and individualized rehabilitation planning	AUC, sensitivity, specificity, PPV, NPV, C-index	ML models showed excellent predictive performance (AUC up to 0.947) and accurately predicted independent walking outcomes	Lack of external validation, overestimation in ages 5–6, possible misclassification bias, usability not evaluated
Tsai, Lee & Huang <sup>[8]</sup>	Research on Applying Deep Learning to Visual-Motor Integration Assessment Systems in Pediatric Rehabilitation Medicine	Algorithms	Taiwan	Observational, Validation	University OT department / laboratory setting	3–6 years	Typically developing preschool children	8610 images (from multiple children)	To develop and optimize a deep learning-based automated scoring system for VMI assessment	Deep Learning / Computer Vision	Improved DenseNet201 (CNN)	Scanned drawings (Image)	Motor, Sensory-perceptual,	Beery-Buktenica VMI	Automated scoring of visual-motor integration performance	Large-scale screening and clinical assessment support	Accuracy, training time, validation loss	The improved DenseNet model achieved up to 95.13% accuracy (6 items) and 89.84% (12 items), outperforming previous CNN-based methods	Imbalanced dataset, limited real-world testing, lack of external validation, dependence on high-quality scanned images
Hegde et al. <sup>[41]</sup>	The Pediatric SmartShoe: Wearable Sensor System for Ambulatory Monitoring of Physical Activity and Gait	IEEE Transactions on Neural Systems and Rehabilitation Engineering	USA	Observational, Feasibility	Community and laboratory environment	5–15 years	Typically developing children and children with cerebral palsy	21 (10 CP, 11 TD)	To develop and test a wearable shoe-based sensor system for monitoring gait and physical activity in children	Machine Learning	SVM, Random Forest, Naïve Bayes (leave-one-out cross-validation, activity classification)	Wearable sensors (pressure sensors, accelerometers)	Motor, Sensory-perceptual	SmartShoe system, gait protocol	Continuous monitoring of gait and activity patterns	Long-term monitoring and functional assessment	Accuracy, step detection rate, activity classification rate	The SmartShoe system reliably detected walking, running, and activity levels in natural environments	Small sample size, early prototype limitations, limited battery life, outdated hardware
Rico-Olarte et al. <sup>[42]</sup>	Towards Classifying Cognitive Performance by Sensing Electrodermal Activity in Children With Specific Learning Disorders	IEEE Access	Colombia, Germany	Quasi-experimental, Observational, Validation	Clinical rehabilitation center using HapHop-Physio	7–11 years	Children with specific learning disorders (SLD)	14 children (945 data samples from 130 sessions)	To classify cognitive performance using electrodermal activity signals collected during rehabilitation therapy	Machine Learning	Random Forest, MLP, SVM, KNN, J48, Naïve Bayes, Boosting, Bagging	Wearable sensors (EDA – E4 wristband), Audio recordings	Cognitive	ENI battery, HapHop-Physio performance scores	Objective monitoring of cognitive performance during therapy	Progress monitoring and rehabilitation support	Accuracy, recall, AUC, Kappa	The Random Forest model achieved the best performance with 79.95% accuracy and substantial agreement (κ = 0.698), demonstrating feasibility for cognitive monitoring	Small sample size, subject-dependent models, signal noise, lack of independent validation
Schafmeyer et al. <sup>[43]</sup>	Using artificial intelligence-based technologies to detect clinically relevant changes of gross motor function in children with cerebral palsy	Developmental Medicine & Child Neurology	Germany	Retrospective cohort, Observational, Validation	Single-center rehabilitation program ("Auf die Beine")	<18 years (mean age 6 years 4 months)	Children with cerebral palsy	1352 paired assessments	To compare GMFM-66 and rGMFM-66 for detecting clinically relevant changes in gross motor function	Machine Learning	Random Forest, Feedforward Neural Network, SVM	Clinical assessment records	Motor	rGMFM-66, GMFCS	Detection of meaningful change in motor function	Outcome monitoring and therapy evaluation	AUC, sensitivity, specificity, PPV, NPV, correlation	rGMFM-66 showed high agreement with GMFM-66 (r = 0.99) and excellent accuracy for detecting improvement and deterioration	Retrospective design, underrepresentation of GMFCS I and V, lack of prospective validation
Ienaga et al. <sup>[44]</sup>	Development and Verification of Postural Control Assessment Using Deep-Learning-Based Pose Estimators: Towards Clinical Applications	Occupational Therapy International	Japan	Observational, Validation	Laboratory and preschool settings	3–6 years (children); 19–35 years (adults)	Typically developing children and adults	57 (34 children, 23 adults)	To compare pose estimation algorithms and develop quantitative indices reflecting OT evaluations of postural control	Deep Learning / Computer Vision	MediaPipe Pose, OpenPose, AlphaPose (CNN-based pose estimators)	Video	Motor, Sensory-perceptual	JPAN "One Arm and One Leg Balance" task, Likert OT ratings	Automated quantification of postural control	Clinical screening and detailed motor assessment	AIC, adjusted R <sup>2</sup> , regression coefficients, processing time, estimation error	MediaPipe Pose showed optimal balance of accuracy and speed; SPB1, SPB3, and AG2 best reflected therapists' evaluations	Limited clinical populations, inclusion of adults, task-specific validation, lack of large clinical samples
Villegas-Ch et al. <sup>[45]</sup>	Detection of Abnormal Patterns in Children's Handwriting by Using an Artificial-Intelligence-Based Method	Informatics	Ecuador	Observational, Validation	Primary school classrooms	7–11 years	Typically developing school-aged children	71 children (210 samples)	To develop and evaluate an AI system for detecting abnormal handwriting patterns	Machine Learning / Computer Vision	DTW (PyDTW), Feature-based classifiers, OpenCV processing	Video, Image, Tablet data	Motor, sensory-perceptual	Camera-based handwriting analysis system, digital tablet	Early detection of handwriting difficulties	Screening and referral support for OT and special education	Accuracy, sensitivity, specificity, F1-score, AUC, PPV, NPV	The system achieved 92% accuracy and 95% detection agreement with camera data; 36 children were referred for intervention and showed 30% improvement	Limited sample size, indirect measurement of muscle tension, sensitivity to environmental factors, lack of standardized validation

AIMS: Alberta Infant Motor Scale; AI: Artificial intelligence; SVR: Support vector regression; ICC: Intraclass correlation coefficient; EHR: Electronic health records; ADL: Activities of daily living; MAE: Mean absolute error; RMSE: Root mean squared error; ML: Machine learning; RCT: Randomized controlled trial; DCD: Developmental coordination disorder; TD: Typically developing; CNN: Convolutional neural networks; MLP: Multilayer perceptron; SVM: Support vector machine; KNN: K-Nearest neighbors; PPV: Positive predictive value; NPV: Negative predictive value; AUC: Area under the curve; ROC: Receiver operating characteristic; RF: Random forest; ITW: Idiopathic toe walking; SLD: Specific learning disorders; OT: Occupational therapy; K-NNR: K-Nearest neighbors regression; IMU: Inertial measurement unit; GBM: Gradient boosting machine; GMFCS: Gross motor function classification system; GMFM-66: Gross motor function measure-66; CV: Computer vision.



## SYSTEMATIC REVIEW

# A Systematic Review and Meta-Analysis of Diagnostic Performance Comparison between DeepSeek and Physicians

Jianwen Zeng<sup>1</sup>, Xule Zhu<sup>1</sup>, Xin Liu<sup>1</sup>, Shiyong Shen<sup>1</sup>, Sixie Li<sup>1</sup>, Shihua Cao<sup>1,2</sup>

<sup>1</sup>School of Public Health and Nursing, Hangzhou Normal University, Hangzhou, China

<sup>2</sup>Key Engineering Research Center of Mobile Health Management System, Ministry of Education, Hangzhou, China

## Abstract

**Introduction:** Since the release of DeepSeek, it has attracted substantial global attention and has increasingly been explored as a tool for medical diagnosis, showing promising potential for clinical applications. To comprehensively evaluate the effectiveness, potential, and limitations of DeepSeek in medical diagnosis, thereby informing future research and real-world implementation and supporting the development of AI-assisted diagnostic care.

**Methods:** We searched Web of Science Core Collection, Embase, MEDLINE, Scopus, IEEE Xplore, and medRxiv from inception to August 8, 2025. Two authors independently screened studies, extracted data according to predefined inclusion and exclusion criteria, and assessed study quality using the Prediction model Risk of Bias Assessment Tool.

**Results:** Twenty-four studies were included, evaluating 6 DeepSeek model variants; DeepSeek-R1 was the most frequently assessed. Quality appraisal indicated a high risk of bias in 13 studies (54%). DeepSeek's performance varying across medical specialties. Overall performance did not differ significantly between DeepSeek and physicians ( $p=0.07$ ); however, DeepSeek did not reach physician-level performance, with diagnostic accuracy 7.7% points lower than physicians.

**Discussion and Conclusion:** DeepSeek demonstrated no statistically significant difference compared with physicians, yet it remained below physician performance. At present, it should not replace expert clinicians. Nevertheless, it may serve as a valuable adjunct in non-specialist settings and as an educational tool for medical trainees.

**Keywords:** DeepSeek; Diagnosis; Large language model; Meta-analysis; Systematic review

Large language models (LLMs) have revolutionized the field of artificial intelligence (AI) by demonstrating impressive capabilities in natural language understanding and reasoning. These models are rapidly becoming transformative tools in the medical field, showing potential across various clinical applications, including personalized health consultations, research, clinical decision support, surgical planning

assistance, and telemedicine promotion.<sup>[1]</sup> Their ability to process and understand complex medical information presents opportunities for improving clinical decision-making, automating administrative tasks, and enhancing patient care.<sup>[2–4]</sup> As AI technology matures, these models are expected to become valuable aids in navigating the expanding domain of medical knowledge and improving healthcare services.

**Cite this article as:** Zeng J, Zhu X, Liu X, Shen S, Li S, Cao S. A Systematic Review and Meta-Analysis of Diagnostic Performance Comparison between DeepSeek and Physicians. *Lokman Hekim Health Sci* 2026;6(2):323–333.

**Correspondence:** Shihua Cao, M.D. School of Public Health and Nursing, Hangzhou Normal University, Hangzhou, China

**E-mail:** csh@hznu.edu.cn **Submitted:** 08.05.2026 **Revised:** 15.05.2026 **Accepted:** 16.05.2026 **Available Online:** 11.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



The application of LLMs in medicine has led to growing research attention toward their diagnostic capabilities. Studies have extensively explored these models' performance in interpreting clinical data, understanding patient histories, and even suggesting potential diagnoses.<sup>[5,6]</sup> Medical diagnosis is a challenging task, but with comprehensive medical knowledge, these models serve as diagnostic support tools through natural language interaction,<sup>[7]</sup> showing promising potential even for diagnosing complex clinical cases.<sup>[8]</sup> LLMs' accuracy, speed, and efficiency in processing vast amounts of medical literature and patient information underscore their value as tools in medical diagnostics.

In late 2024, DeepSeek introduced two open-source models, DeepSeek-V3 and DeepSeek-R1,<sup>[9,10]</sup> which quickly garnered global attention. These models slightly outperform generative pre-trained transformer (GPT)-4o and GPT-o1 in performance while reducing computational costs by an order of magnitude.<sup>[11]</sup> Furthermore, the open nature of the DeepSeek models has fostered the development of a collaborative ecosystem, enabling researchers and developers worldwide to experiment with, refine, and adapt these models for various applications. This collective effort has accelerated DeepSeek's adoption in the healthcare sector.<sup>[12]</sup> The scalable DeepSeek-R1 architecture (1.5B–671B parameters) has driven the development of medical LLMs,<sup>[13–15]</sup> prompting investigations into whether the model can compete with proprietary models in clinical decision tasks, including medical diagnosis, and whether enhanced reasoning abilities can benefit clinical workflows. Despite the growing body of research on DeepSeek's use in medical diagnosis, a significant gap remains in the literature: the lack of a comprehensive meta-analysis of the model's diagnostic capabilities and comparisons with physician performance. Such comparisons are crucial for understanding the real-world significance and effectiveness of DeepSeek in clinical settings. For other models, such as ChatGPT, LLaMA, and Gemini, comprehensive analyses of their performance in medical diagnostics have been conducted.<sup>[16]</sup> Therefore, there is a need for a systematic review and meta-analysis of DeepSeek's diagnostic performance to draw more reliable conclusions.

This study aims to fill this gap by systematically evaluating the diagnostic capabilities of DeepSeek in the medical field. Our focus is to provide a comprehensive assessment of DeepSeek's diagnostic performance and compare it with physician performance. By synthesizing results from various studies, we seek to gain deeper insights into DeepSeek's

effectiveness, potential, and limitations in medical diagnosis. This analysis aims to provide a foundational reference for future research and practical applications in the field, ultimately advancing the development of AI-assisted diagnostics in healthcare.

## Methods

### Protocol and Registration

This systematic review was prospectively registered with PROSPERO (CRD420251125959). Our study adhered to the relevant sections of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for diagnostic accuracy studies (Supplementary Material 1 S3).<sup>[17,18]</sup> All stages of the review (title and abstract screening, full-text screening, data extraction, and bias assessment) were independently conducted by two reviewers (JZ and WQ), with discrepancies resolved through discussion with a third independent reviewer (SC).

### Search Strategy and Study Selection

We conducted searches in the Web of Science Core Collection, Scopus, Embase, Medline, IEEE Xplore, and Medrxiv to identify studies validating DeepSeek for diagnostic tasks. A search strategy was developed, incorporating variants of DeepSeek and diagnostic-related terms, with the full search strategy provided in Supplementary Material 1 S1. Two authors (JZ and WQ) independently screened the titles and abstracts of the retrieved studies using the search strategy to identify studies meeting the inclusion and exclusion criteria (Textbox 1). Full-text assessments were then performed, with disagreements resolved through discussion with a third author. The search covered the period from database inception to August 8, 2025. To ensure comprehensiveness, we also reviewed the reference lists of relevant studies and citations. Literature management and duplicate removal were conducted using EndNote software.

### Data Extraction

Before full-text screening, title and abstract screening were performed. A data extraction form was created using Microsoft Excel for data extraction, which was independently conducted by two reviewers. Any discrepancies between the reviewers were resolved through discussion. Information was extracted from each study, including the first author, model and version, model task, type of test dataset (internal, external, or unknown),<sup>[19]</sup> medical specialty, accuracy, sample size, and publication

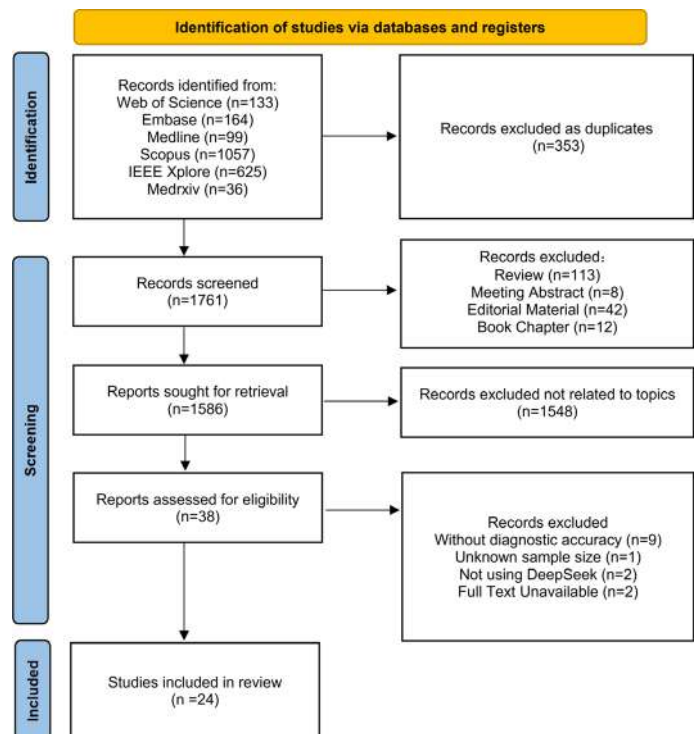
status (preprint or peer-reviewed) for the meta-analysis of DeepSeek performance. Based on the relationship between the model's training and test data, three types of tests were defined.<sup>[20]</sup> Internal tests were defined as cases where the test data came from the same source or distribution as the training data but was appropriately separated from the training set using standard methods such as cross-validation or random splitting. External tests were defined as cases where the test data were collected after the training data cutoff, or when the model was tested on private data. Unknown tests were defined as cases where the test data were collected before the training data cutoff, and the data were publicly available. This distinction was made because the complete training datasets for the companies developing these models have not been made public. Additionally, when both model and physician diagnostic performance were presented in the same paper, both were extracted for meta-analysis. When a single model used multiple prompts and individual performances were available for each, the original prompt's performance was selected over the performance derived from iterative prompts. When multiple languages were used, the diagnostic performance based on the language of the test questions was selected.

### Quality Assessment

We used Prediction model Risk of Bias Assessment Tool (PROBAST) to assess the risk of bias and applicability of the studies.<sup>[19]</sup> This tool utilizes signaling questions across four domains (participants, predictors, outcomes, and analysis) to provide both an overall and detailed assessment. Some PROBAST signaling questions were not included, as they are not relevant to generative AI models. We made modifications to PROBAST based on the study by Takita et al.,<sup>[16]</sup> with details of the modifications provided in Supplementary Material 1 S4.

### Statistical Analysis

Heterogeneity was assessed using the I<sup>2</sup> statistic. A random-effects model was applied when the I<sup>2</sup> value exceeded 50%; otherwise, a fixed-effects model was used. Diagnostic accuracy was reported with 95% confidence intervals (CIs). Statistical significance was set at  $p < 0.05$ . The pooled accuracy values for DeepSeek and physicians were calculated, and the overall accuracy of the model was compared with that of physicians. Subgroup analyses were also conducted to compare the model's performance across different specialties. To evaluate the impact of overall risk of bias, we performed a subgroup analysis limited to studies



**Figure 1.** Literature screening process.

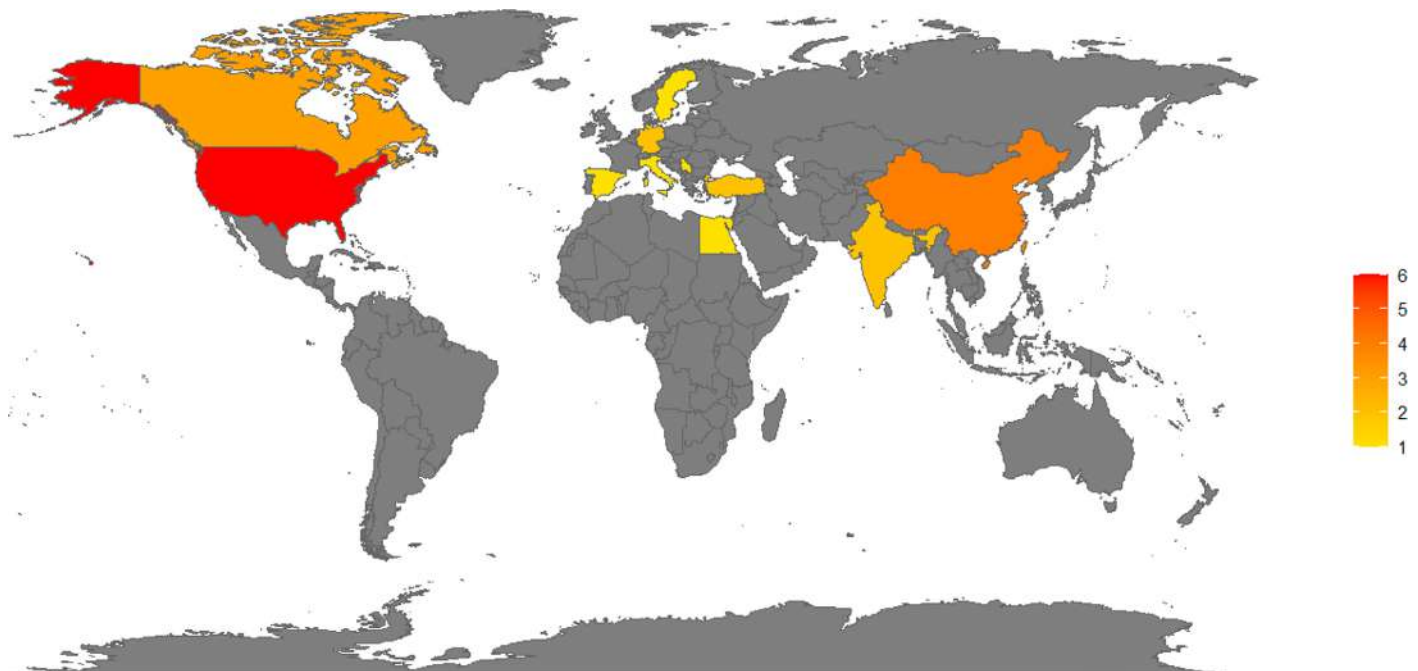
with low bias risk. Publication bias was assessed using funnel plots and Egger's regression test to evaluate its effect on the comparison of diagnostic performance between the model and physicians. Additionally, heterogeneity was analyzed in the full dataset and the low-bias-risk subgroup to assess its impact. All statistical analyses were conducted using R version 4.4.0.

## Results

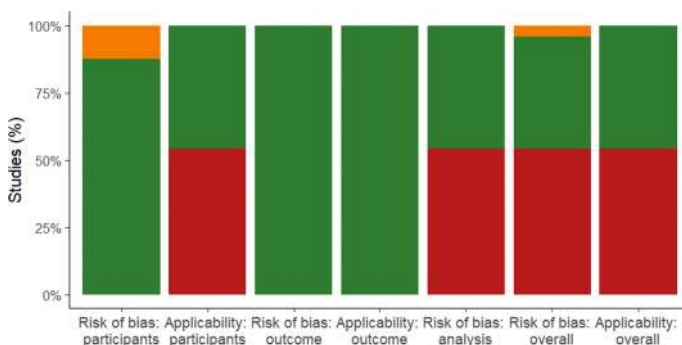
### Study Selection and Characteristics

A total of 2114 studies were identified across six databases, of which 353 were duplicates. After excluding reviews, conference abstracts, editorials, and book chapters, the titles and abstracts of 1586 studies were screened. From these, 1546 studies were excluded, and 39 studies underwent full-text review. Nine studies were excluded for lacking diagnostic accuracy, two for not using DeepSeek, two for inability to access the full text, and one for lacking sample size, resulting in the inclusion of 24 studies.<sup>[21–44]</sup> The detailed screening process is shown in Figure 1.

Among the 24 studies, all were published in 2025 and involved 11 countries or regions (Fig. 2). The United States contributed the most studies (6), followed by China (4) and Canada (3). The most frequently evaluated model was DeepSeek-R1 (16 studies), which includes other parameters or versions such as DeepSeek-R1-70B and



**Figure 2.** Distribution by country or region.



**Figure 3.** Summary of Prediction Model Study Risk of Bias Assessment Tool risk of bias.

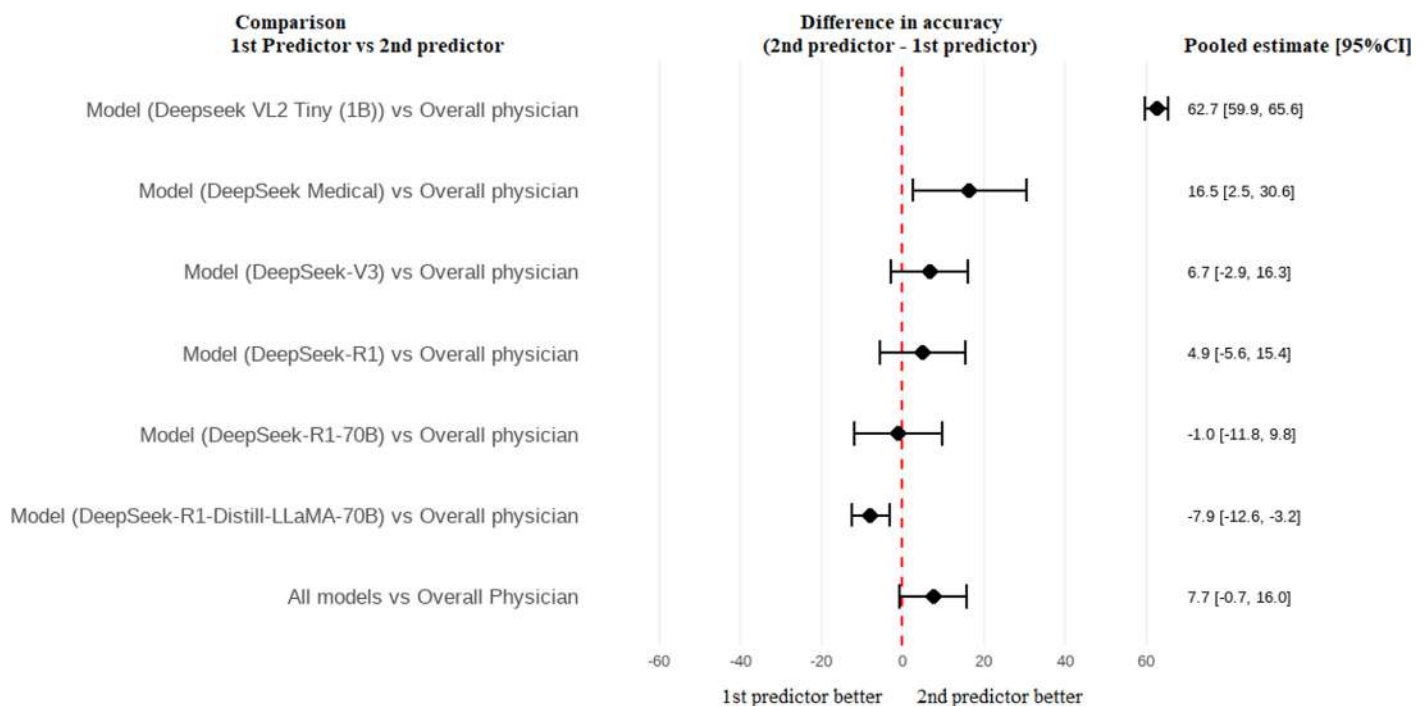
DeepSeek-R1-Distill-LLaMA-70B; these were grouped under the DeepSeek-R1 category. The next most evaluated model was DeepSeek-V3 (6 studies), while DeepSeek Medical and DeepSeek VL2 Tiny (1B) were each represented by one study. Detailed information on each model can be found in Appendix A: Supplementary Material 1 S2.

This review covers a wide range of medical specialties, with ophthalmology (5 studies) and general medicine (4 studies) being the most common. Other specialties represented include oral medicine (3), pediatrics (2), and emergency medicine (2). Additionally, there was representation from gastroenterology, oncology, neurology, sleep medicine, radiology, otolaryngology, critical care medicine, and rheumatology, with one study each. Regarding model tasks, free-text tasks were the most common (20 studies), followed by classification tasks (4 studies). In terms of test dataset

type, 11 studies involved external testing, while 13 studies had unknown test dataset types due to the unavailability of training data for generative AI models. Among the included studies, 17 were peer-reviewed articles, 6 were preprints, and 1 was a letter. The study characteristics are summarized in Table 1 and the Supplementary Material 2. Nine studies compared DeepSeek's performance with that of physicians. [21,25,27–30,39,40,44] Of these, 6 studies used DeepSeek-R1, 2 used DeepSeek-V3, and 1 used DeepSeek Medical.

### Quality Assessment

The risk of bias assessment using the PROBAST indicated that participant and outcome adjudication were generally associated with a low risk of bias, while analysis was associated with a high risk of bias. Overall, 13 studies (54%) were assessed as having a high risk of bias, 10 studies (42%) were assessed as having a low risk of bias, and 1 study had an unknown risk of bias. In terms of applicability, there were high concerns regarding the applicability of participants, while there were low concerns regarding the applicability of outcomes. In total, 13 studies (54%) were rated as having high concerns regarding generalizability, and 11 studies (46%) were rated as having low concerns regarding generalizability<sup>[19]</sup> (Fig. 3). The primary factors contributing to these concerns were studies evaluating models with small test datasets and research where external validation could not be substantiated due to the unknown training data of generative AI models. Detailed results are provided in Supplementary Material 2.



**Figure 4.** Comparison results between models and physicians.

## Meta-Analysis

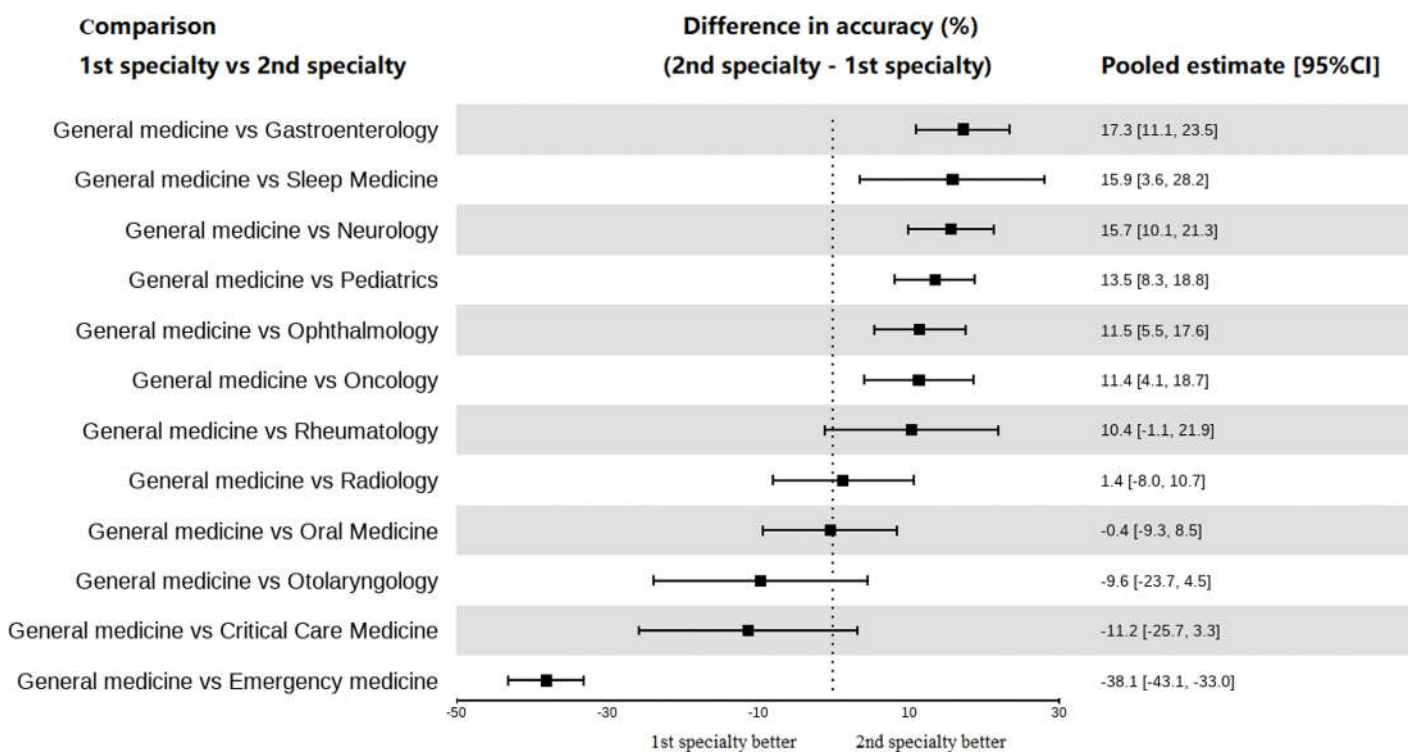
The comparison results between the model and physicians are shown in Figure 4. Meta-analysis revealed no significant difference between DeepSeek and physicians overall (physician accuracy was 7.7% higher [95% CI: -0.7–16.0%],  $p=0.07$ ). DeepSeek VL2 Tiny (1B) exhibited significantly lower diagnostic performance than physicians ( $p<0.001$ ), and DeepSeek Medical also performed worse than physicians ( $p<0.05$ ). Although DeepSeek-V3, DeepSeek-R1, and DeepSeek-R1-70B showed no significant differences compared to physicians, DeepSeek-V3 and DeepSeek-R1 had lower diagnostic accuracy, while DeepSeek-R1-70B slightly outperformed physicians. Interestingly, DeepSeek-R1-Distill-LLaMA-70B performed significantly better than physicians.

In our meta-regression (Fig. 5), we found no significant differences in model performance between general medicine and specialties such as rheumatology, radiology, oral medicine, otolaryngology, and critical care medicine. Significant differences were observed when comparing model performance with specialties like gastroenterology, neurology, pediatrics, and ophthalmology ( $p<0.001$ ). Model performance also showed some significance when compared with sleep medicine and oncology ( $p<0.05$ ), with general medicine performing worse than these specialties. In contrast, a significant difference was observed when compared with emergency medicine ( $p<0.001$ ), but

the model performed significantly better in general medicine than in emergency medicine. In the low-bias risk subgroup analysis, DeepSeek's performance overall did not significantly differ from that of physicians ( $p=0.514$ ). Meta-regression analysis of bias risk and publication status revealed no significant differences for bias risk ( $p=0.708$ ) or publication status ( $p=0.214$ ). In the heterogeneity analysis, the  $R^2$  value (representing the degree of explained heterogeneity) was 45.4% for all studies and 18.5% for low-bias risk studies, indicating a moderately low level of explainable variation. We assessed publication bias through regression analysis and quantified the asymmetry of the funnel plot (Supplementary Material 1 S5). The results indicated no significant risk of publication bias ( $p=0.200$ ).

## Discussion

In this systematic review and meta-analysis, we analyzed the diagnostic performance of DeepSeek and physicians. To our knowledge, this is the first analysis comparing the performance of all versions of DeepSeek in medical diagnosis, alongside physicians. Initially, we identified 2114 studies, and 24 studies were ultimately included for the systematic review and meta-analysis. This study covers multiple versions of the DeepSeek model and a wide range of medical specialties, with DeepSeek-R1 being the most frequently evaluated model and Ophthalmology being the most commonly represented medical specialty. Quality



**Figure 5.** DeepSeek performance among specialties.

assessment revealed that most studies had a high risk of bias. Meta-analysis showed that the average diagnostic accuracy of the DeepSeek models was 76% (95% CI: 67–83%). Some models performed similarly to physicians, with no significant performance differences (accuracy difference: 7.7% [95% CI: -0.7–16.0%];  $p=0.07$ ). Furthermore, our analysis also found that DeepSeek's performance varied across most medical fields. This comprehensive study highlights the varied capabilities and limitations of generative AI in medical diagnostics.

LLMs have the potential to improve patient care by enhancing clinical decision-making and automating administrative tasks.<sup>[45]</sup> However, adhering to data privacy regulations and medical device compliance poses challenges for proprietary LLM implementations.<sup>[46,47]</sup> Open-source LLMs present a potential alternative for clinical applications. The release of the open-source DeepSeek model has garnered widespread attention.<sup>[48,49]</sup> including in the medical field. The meta-analysis of DeepSeek models in healthcare provides critical insights into its clinical implications. With an accuracy of 76%, DeepSeek demonstrates near-excellent performance, confirming its substantial potential for application in certain clinical scenarios. Some models, including DeepSeek-V3, DeepSeek-R1, and DeepSeek-R1-70B, showed comparable performance to physicians, and

even DeepSeek-R1-Distill-LLaMA-70B outperformed physicians. Although the analysis of DeepSeek-R1-70B and DeepSeek-R1-Distill-LLaMA-70B is based on a single study, it is undeniable that DeepSeek can assist in the provision of medical services or serve as a preliminary diagnostic tool in resource-limited environments.<sup>[50]</sup> Furthermore, compared to proprietary models, DeepSeek offers local deployment capabilities and customization potential, making it highly suitable for various medical research and clinical tasks.<sup>[51]</sup> It is expected to be deployed at scale in hospitals.<sup>[52]</sup> This large-scale application marks the expansion of AI beyond diagnostic support to areas such as hospital management, research facilitation, and patient management.<sup>[53]</sup>

There were notable differences in the performance of DeepSeek across various medical specialties. However, most specialties included only one study, which limits the generalizability of these findings and calls for further investigation. Additionally, among the models included, all except DeepSeek VL2 Tiny (1B) do not support direct image analysis or interpretation. Consequently, image-related tasks were typically excluded during model testing or converted into text for testing purposes. This limitation may restrict the use of DeepSeek in specialties such as Ophthalmology and Oral Medicine, but if users describe visual features of conditions in text form—such as appearance, location, color, associated

**Table 1.** Study characteristics

Citation	First author	Model	Model task	Test type	Specialty	Comparison group	Eligible	Preprint	Overall ROB	Overall applicability
[21]	Aminan	DeepSeek-R1	Free text	Unknown	Ophthalmology	Physicians	11	Preprint	High	Low
[26]	Hoyt	DeepSeek-R1	Free text	Unknown	General medicine	NA	162	Preprint	Low	Low
[27]	Hussain	DeepSeek-V3	Free text	External	Ophthalmology	Physicians	53	Preprint	High	High
[36]	Mondillo	DeepSeek-R1	Choice	Unknown	Pediatrics	NA	500	Preprint	Low	Low
[37]	Mruthyunjaya	DeepSeek-R1-70B	Free text	External	Rheumatology	NA	50	Preprint	High	Low
[38]	Naliyattthaliyazchayil	DeepSeek-R1	Free text	External	Emergency medicine	NA	300	Preprint	Low	High
[22]	Chan	DeepSeek-R1	Free text	Unknown	General medicine	NA	100	Peer-reviewed	Low	High
[24]	Diniz-Freitas	DeepSeek-R1	Free text	Unknown	Oral medicine	NA	36	Peer-reviewed	High	High
[25]	Hassanein	DeepSeek-V3	Free text	Unknown	Oral medicine	Physicians	80	Peer-reviewed	High	High
[28]	Ilic	DeepSeek Medical	Free text	External	Pediatrics	Physicians	45	Peer-reviewed	High	High
[29]	Cheng	DeepSeek-R1	Free text	Unknown	Ophthalmology	Physicians	20	Peer-reviewed	High	Low
[30]	Kang	DeepSeek-R1	Free text	External	Oncology	Physicians	159	Peer-reviewed	Low	High
[31]	Kaygisiz	DeepSeek-V3	Free text	External	Oral medicine	NA	16	Peer-reviewed	High	High
[32]	Kurz	Deepseek-VL2	Choice	Unknown	Emergency medicine	NA	1012	Peer-reviewed	Low	Low
[34]	Mikhail	DeepSeek-R1	Choice	Unknown	Ophthalmology	NA	158	Peer-reviewed	Low	Low
[35]	Moël	DeepSeek-R1	Choice	Unknown	General medicine	NA	100	Peer-reviewed	Low	Low
[39]	Pan	DeepSeek-R1	Free text	External	Neurology	Physicians	338	Peer-reviewed	Low	High
[40]	Patel	DeepSeek-R1	Free text	Unknown	Sleep medicine	Physicians	32	Peer-reviewed	High	Low
[41]	Spitzl	DeepSeek-V3	Free text	External	Radiology	NA	111	Peer-reviewed	Unclear	High
[42]	Tordjiman	DeepSeek-R1	Free text	Unknown	General medicine	NA	50	Peer-reviewed	High	High
[43]	Camalan	DeepSeek-V3	Free text	External	Otolaryngology	NA	50	Peer-reviewed	High	High
[44]	Wu	DeepSeek-R1	Free text	External	Critical care medicine	Physicians	48	Peer-reviewed	High	Low
[33]	Li	DeepSeek-V3	Free text	Unknown	Ophthalmology	NA	22	Letter	High	Low
[23]	Dai	DeepSeek-R1-Distill	Free text	External	Gastroenterology	NA	198	Peer-reviewed	Low	High

ROB: Risk of bias.

symptoms, and duration—DeepSeek may still provide relevant information,<sup>[24]</sup> thus reducing this limitation. However, studies have found that despite ChatGPT-4o's ability to directly analyze images, adding images to case descriptions did not improve diagnostic accuracy.<sup>[54]</sup> While it is unclear whether the lack of image analysis capabilities affects DeepSeek's diagnostic accuracy, its technical limitations in multimodal data integration hinder its ability to provide comprehensive diagnostic and treatment recommendations. A potential solution to this issue is the implementation of hierarchical attention mechanisms (e.g., cross-modal attention weight allocation between images and text), combined with time-series modeling to capture the dynamic features of disease progression.<sup>[55]</sup>

Research comparing DeepSeek and physician performance also offers an intriguing perspective in the context of medical education.<sup>[56]</sup> Currently, DeepSeek's performance is comparable to that of physicians, and the model's ability to be deployed locally and its cost-effectiveness<sup>[13]</sup> suggest opportunities for integrating it into medical training. Since studies comparing models to physicians do not specify the physicians' experience levels, we assume the doctors involved in the included studies are non-experts, making it impossible to conduct further comparisons between expert and non-expert physicians and the model. However, DeepSeek can be used as an educational tool for medical students or residents, especially when simulating non-expert scenarios, where its performance is on par with healthcare professionals.<sup>[57]</sup> This integration can enhance the learning experience, especially when DeepSeek is locally deployed, allowing training with complex or rare cases using local data, providing diverse clinical case studies, and offering targeted instruction to students or residents, while also promoting self-assessment and feedback. Although the model's performance does not significantly differ from physicians, its diagnostic accuracy has yet to reach that of physicians, emphasizing the irreplaceable value of human judgment and experience in medical decision-making. One key advantage of models like DeepSeek-R1 is its reasoning-centered design, making the decision-making process more interpretable and transparent. However, this feature does not eliminate the fundamental issue of AI-generated "hallucinations," where models generate seemingly reasonable but actually erroneous medical information, which poses significant risks in patient care, especially when AI output influences diagnosis, treatment recommendations, or research conclusions.<sup>[58]</sup> Furthermore, DeepSeek-R1's self-reflection capability could be used to bypass its safety constraints,

raising concerns about models generating diagnoses that deviate from established medical guidelines or rationalizing incorrect treatments.<sup>[59]</sup> Therefore, external safeguards, such as rule-based reinforcement filters, human-in-the-loop validation, and continuous real-world verification, are crucial for safe deployment.<sup>[60]</sup> In this context, we propose a hybrid system where physicians ask the model for differential diagnoses of clinical conditions, which may include diagnoses not previously considered by the physician. This collaborative model can leverage the strengths of both parties and provide more comprehensive diagnostics.

To examine the impact of overall bias risk, we conducted a subgroup analysis of studies with low bias risk. The results of studies with low bias risk did not significantly differ from those of the full dataset. Therefore, the higher proportion of studies with high bias risk does not substantively affect our findings or their generalizability. The training data specifics of the DeepSeek series models have not been fully disclosed, but the transparency of training data and its collection period is crucial. Without this transparency, we cannot determine if the test dataset is an external dataset, which could introduce bias. Transparency ensures the fairness, interpretability, and scientific integrity of the model, helps identify potential biases, and promotes independent replication and validation. As AI technology continues to evolve, data transparency will become increasingly important. Thus, establishing strict data collection and management standards to ensure transparency and fairness should be one of the core tasks in LLMs research and application.

## Study Limitations

Although the methodology of this study is comprehensive, it has certain limitations. The generalizability of our findings requires careful consideration. Heterogeneity analysis revealed that the explained variability was at a moderately low level, indicating that our meta-regression model did not fully account for the differences between studies. Other factors not addressed in the analysis may influence DeepSeek's performance, including case complexity, model hyperparameter settings, and prompt design. Furthermore, the included studies did not clearly specify the languages used, but it can be confirmed that all were non-Chinese, with only one study utilizing bilingual input (Chinese and English).<sup>[39]</sup> The impact of using native languages (e.g., Chinese) on the model's performance remains uncertain. Finally, we did not compare DeepSeek with other models, and further research is needed to extend this analysis.

## Conclusion

This meta-analysis provides a detailed understanding of the capabilities and limitations of DeepSeek in medical diagnosis. With an overall diagnostic accuracy of 76%, DeepSeek shows no significant difference in performance compared to physicians. However, it has not yet reached the level of physicians and cannot replace expert clinicians. It may, however, serve as a valuable auxiliary tool in non-specialized settings and as an educational tool for medical students. Furthermore, it is important to emphasize the need for continued advancement and specialization in model development, as well as the necessity for rigorous, external validation studies to address the widespread high risk of bias. The transparency of training data should also be prioritized to ensure that DeepSeek can effectively integrate into clinical practice.

**Ethics Committee Approval:** Ethical approval was not required for this study since this is a review article.

**Conflict of Interest:** The author declare that there is no conflict of interest.

**Financial Disclosure:** The authors declared that this study received no financial support.

**Use of AI for Writing Assistance:** Not declared.

**Authorship Contributions:** JZ, XZ, XL, SS, SC; Design: JZ, XZ, XL, SS, SL; Supervision: JZ, XZ, XL, SS, SC; Data Collection and/or Processing: JZ, XZ, XL, SS, SL; Analysis and/or Interpretation: JZ, XZ, XL, SS, SL; Literature Search: JZ, XZ, XL, SS, SL; Writing: JZ, XZ, XL, SS, SL; Critical Reviews: JZ, XZ, XL, SS, SL.

**Peer-review:** Double blind peer-reviewed.

## References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29(8):1930-40. [\[CrossRef\]](#)
2. Bellini V, Bignami EG. Generative Pre-trained Transformer 4 (GPT-4) in clinical settings. *Lancet Digit Health* 2025;7(1):e6-7. [\[CrossRef\]](#)
3. Boussina A, Krishnamoorthy R, Quintero K, Joshi S, Wardi G, Pour H, et al. Large language models for more efficient reporting of hospital quality measures. *NEJM AI* 2024;1(11). [\[CrossRef\]](#)
4. McCoy TH, Perlis RH. Applying large language models to stratify suicide risk using narrative clinical notes. *J Mood Anxiety Disord* 2025;10:100109. [\[CrossRef\]](#)
5. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA* 2023;330(1):78-80. [\[CrossRef\]](#)
6. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. *Radiology* 2023;308(1):e231040. [\[CrossRef\]](#)
7. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open* 2024;7(10):e2440969. [\[CrossRef\]](#)
8. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nat Commun* 2024;15(1):2050. [\[CrossRef\]](#)
9. Liu A, Feng B, Wang B, Wang B, Liu B, Zhao C, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv* 2024;240504434.
10. Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv* 2025;250112948.
11. Deng Z, Ma W, Han Q-L, Zhou W, Zhu X, Wen S, et al. Exploring DeepSeek: a survey on advances, applications, challenges and future directions. *IEEE-Caa Journal of Automatica Sinica* 2025;12(5):872-93. [\[CrossRef\]](#)
12. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. *Cureus* 2025;17(2):e79221. [\[CrossRef\]](#)
13. Gibney E. Scientists flock to DeepSeek: how they're using the blockbuster AI model. *Nature* 2025. doi: 10.1038/d41586-025-00275-0. Epub ahead of print. PMID: 39881178. [\[CrossRef\]](#)
14. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. *Nature* 2025;638(8049):13-4. [\[CrossRef\]](#)
15. Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. *Nature*. 2025;638(8050):300-1. [\[CrossRef\]](#)
16. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *NPJ Digit Med* 2025;8(1):175. [\[CrossRef\]](#)
17. McInnes MDF, Moher D, Thoms BD, McGrath TA, Bossuyt PM; the PRISMA-DTA Group; et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. *JAMA* 2018;319(4):388-96. [\[CrossRef\]](#)
18. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535. [\[CrossRef\]](#)
19. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170(1):51-8. [\[CrossRef\]](#)
20. Walston SL, Seki H, Takita H, Mitsuyama Y, Sato S, Hagiwara A, et al. Data set terminology of deep learning in medicine: a historical review and recommendation. *Jpn J Radiol* 2024;42(10):1100-9. [\[CrossRef\]](#)
21. Aminan M, Darnell SS, Delsoz M, Nabavi A, Wright C, Jerkins B, et al. GlaucoRAG: A Retrieval-Augmented Large Language Model for Expert-Level Glaucoma Assessment. *medRxiv* 2025;2025.07.03.25330805. [\[CrossRef\]](#)

22. Chan LN, Xu XJ, Lv KY. DeepSeek-R1 and GPT-4 are comparable in a complex diagnostic challenge: a historical control study. *International Journal of Surgery* 2025;111(6):4056-9. [CrossRef]
23. Dai J, Kim MY, Sutton RT, Mitchell JR, Goebel R, Baumgart DC. Comparative analysis of natural language processing methodologies for classifying computed tomography enterography reports in Crohn's disease patients. *NPJ Digit Med* 2025;8(1):324. [CrossRef]
24. Diniz-Freitas M, Diz-Dios P. DeepSeek: Another step forward in the diagnosis of oral lesions. *J Dent Sci* 2025;20(3):1904-7. [CrossRef]
25. Hassanein FEA, El Barbary A, Hussein RR, Ahmed Y, El-Guindy J, Sarhan S, et al. Diagnostic Performance of ChatGPT-4o and DeepSeek-3 Differential Diagnosis of Complex Oral Lesions: A Multimodal Imaging and Case Difficulty Analysis. *Oral Dis* 2025;31(12):3361-71. [CrossRef]
26. Hoyt RE, Knight D, Haider M, Bajwa M. Evaluating a Large Reasoning Model's Performance on Open-Ended Medical Scenarios. medRxiv 2025. DOI: 10.1101/2025.04.29.25326666. Epub ahead of print. [CrossRef]
27. Hussain ZS, Delsoz M, Elahi M, Jerkins B, Kanner E, Wright C, et al. Performance of DeepSeek, Qwen 2.5 MAX, and ChatGPT Assisting in Diagnosis of Corneal Eye Diseases, Glaucoma, and Neuro-Ophthalmology Diseases Based on Clinical Case Reports. medRxiv 2025. doi: 10.1101/2025.03.14.25323836. Epub ahead of print. PMID: 40166547. [CrossRef]
28. Ilić N, Marić N, Cvetković D, Bogosavljević M, Bukara-Radujković G, Krstić J, et al. The artificial intelligence-assisted diagnosis of skeletal dysplasias in pediatric patients: a comparative benchmark study of large language models and a clinical expert group. *Genes (Basel)* 2025;16(7):762. [CrossRef]
29. Jiao C, Rosas E, Asadigandomani H, Delsoz M, Madadi Y, Raja H, et al. Diagnostic performance of publicly available large language models in corneal diseases: a comparison with human specialists. *Diagnostics (Basel)* 2025;15(10):1221. [CrossRef]
30. Kang C, Li J, Yang X, Ren G, Zhang L, Wang W, et al. Performance of large language models in the differential diagnosis of benign and malignant biliary stricture. *Front Oncol* 2025;15:1613818. [CrossRef]
31. Kaygisiz ÖF, Teke MT. Can deepseek and ChatGPT be used in the diagnosis of oral pathologies? *BMC Oral Health* 2025;25(1):638. [CrossRef]
32. Kurz CF, Merzhevich T, Eskofier BM, Kather JN, Gmeiner B. Benchmarking vision-language models for diagnostics in emergency and critical care settings. *NPJ Digit Med* 2025;8(1):423. [CrossRef]
33. Li X, He J, Xie JS, Sharma RA. Comment: Diagnosing Neuro-Ophthalmology Diseases Based on Case Reports: DeepSeek vs ChatGPT. *J Neuroophthalmol* 2025;45(3):e261-2. [CrossRef]
34. Mikhail D, Farah A, Milad J, Nassrallah W, Mihalache A, Milad D, et al. Performance of DeepSeek-R1 in ophthalmology: an evaluation of clinical decision-making and cost-effectiveness. *Br J Ophthalmol* 2025;109(9):976-81. [CrossRef]
35. Moëll B, Sand Aronsson F, Akbar S. Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. *Front Artif Intell* 2025;8:1616145. [CrossRef]
36. Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Masino M. Comparative Evaluation of Advanced AI Reasoning Models in Pediatric Clinical Decision Support: ChatGPT O1 vs. DeepSeek-R1. medRxiv 2025;01.27.25321169. [CrossRef]
37. Mruthyunjaya P, Verma S, Agarwal A, Maharana U, Mandal M, Ahmed S. Right Diagnoses But Wrong Reasoning: Current Large-Language Model-Based Agentic Frameworks Have Flawed Clinical Reasoning Despite High Diagnostic Accuracy. SSRN 2025. doi: 10.2139/ssrn.5339074. Epub ahead of print. [CrossRef]
38. Naliyatthalizaychayil P, Muthyala R, Gichoya JW, Purkayastha S. Evaluating the Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk Stratification: Zero-Shot Prompting Approach. *J Med Internet Res* 2025;27:e74142. [CrossRef]
39. Pan Y, Tian S, Guo J, Cai H, Wan J, Fang C. Clinical feasibility of AI Doctors: Evaluating the replacement potential of large language models in outpatient settings for central nervous system tumors. *Int J Med Inform* 2025;203:106013. [CrossRef]
40. Patel A, Ruoff C, Helgeson SA, Carvalho DZ, Castillo PR, Cheung J. Diagnostic performance of Large Language Models (LLMs) compared with physicians in sleep medicine. *Sleep Med* 2025;134:106677. [CrossRef]
41. Spitzl D, Mergen M, Braren R, Endrös L, Eiber M, Steinhelfer L. LLM-powered breast cancer staging from PET/CT reports: a comparative performance study. *Int J Med Inform* 2025;204:106053. [CrossRef]
42. Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat Med* 2025;31(8):2550-5. [CrossRef]
43. Vural Camalan B, Doluoglu S, Taraf NH, Gunay MM, Ozlugedik S. ChatGPT versus DeepSeek in head and neck cancer staging and treatment planning: guideline-based study. *Eur Arch Otorhinolaryngol* 2025;282(9):4815-24. [CrossRef]
44. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. *Crit Care* 2025;29(1):230. [CrossRef]
45. Quer G, Topol EJ. The potential for large language models to transform cardiovascular medicine. *Lancet Digit Health* 2024;6(10):e767--1. [CrossRef]
46. de Hond A, Leeuwenberg T, Bartels R, van Buchem M, Kant I, Moons KG, van Smeden M. From text to treatment: the crucial role of validation for generative large language models in health care. *Lancet Digit Health* 2024;6(7):e441-3. [CrossRef]
47. Ong JCL, Chang SY-H, William W, Butte AJ, Shah NH, Chew LST, et al. Medical ethics of large language models in medicine. *NEJM AI* 2024;1(7). [CrossRef]
48. Poo M-m. Reflections on DeepSeek's breakthrough. *National Science Review*. 2025;12(3):nwaf044. [CrossRef]
49. Sandmann S, Eils R. Open-source LLM DeepSeek on a par with proprietary models in clinical decision making. *Nature Medicine*. 2025;31:2496-7. [CrossRef]

50. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health* 2018;3(4):e000798. [\[CrossRef\]](#)
51. MohanaSundaram A, Sathanantham ST, Ivanov A, Mofatteh M. DeepSeek's Readiness for Medical Research and Practice: Prospects, Bottlenecks, and Global Regulatory Constraints. *Ann Biomed Eng* 2025;53(7):1754-6. [\[CrossRef\]](#)
52. Ye H. Key Players Overlooked in the Rapid Deployment of DeepSeek To China's Hospitals. *Journal of Medical Systems* 2025;49. [\[CrossRef\]](#)
53. Chen J, Miao C. DeepSeek Deployed in 90 Chinese Tertiary Hospitals: How Artificial Intelligence Is Transforming Clinical Practice. *J Med Syst* 2025;49(1):53. [\[CrossRef\]](#)
54. Diniz-Freitas M, Lago-Méndez L, Limeres-Posse J, Diz-Dios P. Challenging ChatGPT-4V for the Diagnosis of Oral Diseases and Conditions. *Oral Dis* 2025;31(2):701-6. [\[CrossRef\]](#)
55. Liang W, Chen P, Zou X, Lu X, Liu S, Yang J, et al. DeepSeek: the "Watson" to doctors-from assistance to collaboration. *J Thorac Dis* 2025;17(2):1103-5. [\[CrossRef\]](#)
56. Preiksaitis C, Rose C. Opportunities, Challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR Med Educ* 2023;9:e48785. [\[CrossRef\]](#)
57. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med (Lond)* 2023;3(1):141. [\[CrossRef\]](#)
58. Wang X, Zhang NX, He H, Nguyen T, Yu K-H, Deng H, et al. Safety challenges of AI in medicine in the era of large language models. *arXiv* 2024;2409.18968
59. Mercer S, Spillard S, Martin DP. Brief analysis of DeepSeek R1 and its implications for Generative AI. *arXiv* 2025;250202523. [\[CrossRef\]](#)
60. Pantha N, Ramasubramanian M, Gurung I, Maskey M, Ramachandran R. Challenges in guardrailing large language models for science. *arXiv* 2024;241108181.

## S1 Search strategy

<b>Database</b>	<b>Query</b>
<b>WoS CC</b>	(ALL=("DeepSeek")) AND ALL=("medic*" OR "healthcare" OR "diagnosis" OR "diagnostic" OR "diagnose" OR "quiz" OR "exam*" OR "vignette")
<b>Embase</b>	(ALL=('Deepseek')) AND ALL=('medic*' OR 'healthcare' OR 'diagnosis' OR 'diagnostic' OR 'diagnose' OR 'quiz' OR 'exam*' OR 'vignette')
<b>Scopus</b>	(ALL(DeepSeek) AND ALL(medic* OR healthcare OR "diagnosis" OR "diagnostic" OR "diagnose" OR "quiz" OR "exam*" OR "vignette"))
<b>Medline</b>	(TS=("DeepSeek")) AND TS=("medic*" OR "healthcare" OR "diagnosis" OR "diagnostic" OR "diagnose" OR "quiz" OR "exam*" OR "vignette")
<b>MedRxiv</b>	(DeepSeek) AND (medical OR medicine OR healthcare OR diagnosis OR diagnostic OR diagnose OR quiz OR examination OR vignette)
<b>IEEE Xplore</b>	(Full text only=("DeepSeek")) AND Full text only=("medic*" OR "healthcare" OR "diagnosis" OR "diagnostic" OR "diagnose" OR "quiz" OR "exam*" OR "vignette")

## S2 Models' description

---

<b>Model</b>	<b>Description</b>
DeepSeek-V3	Parameters: 671B; Based on a mixture-of-experts (MoE) architecture, supporting up to 128K context length, suitable for mathematical, programming, and Chinese tasks. Its performance rivals that of GPT-4o, with a training cost below 6 million USD.
DeepSeek-R1	Parameters: 671B; Built upon DeepSeek-V3, enhanced with reinforcement learning for reasoning capabilities, supporting chain-of-thought (CoT), excelling in mathematics, programming, and reasoning tasks.
DeepSeek-R1-70B	Parameters: 70B; Distilled from DeepSeek-R1, based on the Llama3.3-70B-Instruct architecture, suitable for reasoning and code generation tasks.
DeepSeek-R1-Distill-LLaMA-70B	Parameters: 70B; Based on the Llama3.3-70B-Instruct architecture, distilled from DeepSeek-R1 for reasoning, applicable to a wide range of natural language processing tasks.
DeepSeek Medical	Parameters: Not disclosed; Fine-tuned from DeepSeek-R1 for the medical domain, supporting diagnostic and clinical decision-making tasks, with chain-of-thought (CoT) and advanced reasoning capabilities.
Deepseek VL2 Tiny (1B)	Parameters: 1B; A vision-language model based on MoE architecture, suitable for visual question answering, document understanding, and visual localization tasks, employing dynamic tiling visual encoding strategies.

---

## S3 PRISMA 2020 checklist

Section and Topic	Item #	Checklist item	Location where item is reported
<b>TITLE</b>			
Title	1	Identify the report as a systematic review.	Title
<b>ABSTRACT</b>			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	
<b>INTRODUCTION</b>			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Introduction
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Introduction
<b>METHODS</b>			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Search strategy and study selection
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Search strategy and study selection
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Supplementary Table 1
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Figure 1
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Protocol and Registration, Data extraction
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Data Extraction, Table1, Supplement ary Material 2
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Data Extraction, Table1, Supplement ary Material 2
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Quality Assessment
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the	Statistical

Section and Topic	Item #	Checklist item	Location where item is reported
		synthesis or presentation of results.	Analysis
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Statistical Analysis
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Statistical Analysis
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Statistical Analysis
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Statistical Analysis
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Statistical Analysis
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Statistical Analysis
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Quality Assessment, Statistical Analysis
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Quality Assessment, Statistical Analysis
<b>RESULTS</b>			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Figure 1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Figure 1
Study characteristics	17	Cite each included study and present its characteristics.	Table 1, Supplement ary Material 2
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Quality Assessment
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Meta analysis
Results of	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing	Meta analysis

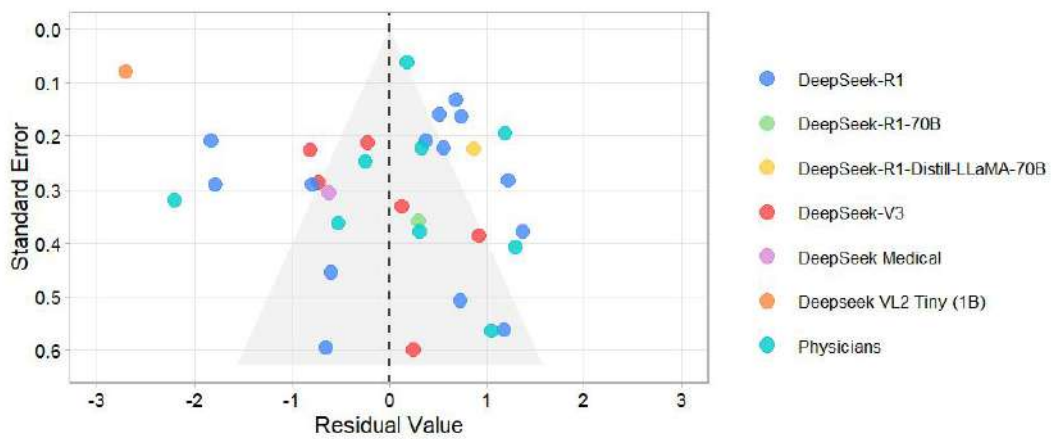
Section and Topic	Item #	Checklist item	Location where item is reported
syntheses		studies.	
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Meta analysis
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Meta analysis
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Meta analysis
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Meta analysis
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Meta analysis
<b>DISCUSSION</b>			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion
	23b	Discuss any limitations of the evidence included in the review.	Discussion
	23c	Discuss any limitations of the review processes used.	Discussion
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion
<b>OTHER INFORMATION</b>			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Protocol and Registration in the Methods
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Protocol and Registration in the Methods
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Protocol and Registration in the Methods
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Acknowledgements
Competing interests	26	Declare any competing interests of review authors.	Competing Interests
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data availability

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. doi: 10.1136/bmj.n71

## S4 PROBABST modifications

PROBABST Items	Modifications
Domain 1: Participants	No changes (Refer to participant data for diagnosis)
Domain 2: Predictors	N/A—removed from scoring
Domain 3: Outcome	Items 3.3, 3.5, and 3.6 N/A
Domain 4: Analysis	Items 4.5, 4.6, and 4.9 N/A
Domain 5: Overall	No changes

## S5 Funnel plot.

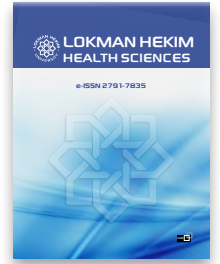


The funnel plot illustrates the distribution of the residuals of the fitted values corresponding to their standard errors in the meta-regression. The Egger test results z-value = 1.3085 and p = 0.200, which indicates no significant publication bias.

## Supplementary material 2

Year	Study	Model	Version	Database	Referebce standard	ROB of Participants	Applicability of participants	ROB of Outcome	Applicability of outcome	ROB of analysis
2025	Aminan[21]	DeepSeek-R1	Not written	The case reports from EyeRounds	Expert consensus	Low	Low	Low	Low	High
2025	Hoyt[26]	DeepSeek-R1	Not written	Medical scenarios from MMLU-Pro question bank	Answer	Low	Low	Low	Low	Low
2025	Hussain[27]	DeepSeek V3	Not written	Case reports from university-accessible databases		Low	High	Low	Low	High
2025	Mondillo[36]	DeepSeek-R1	Not written	Questions from the MedQA dataset	Answer	Low	Low	Low	Low	Low
2025	Mruthyunjaya[37]	DeepSeek-R1-70B	Not written	Cases in knowledge base	Expert consensus	Low	Low	Low	Low	High
2025	Naliyatthaliyazchayil[48]	DeepSeek-R1	Not written	Patient cases in MIMIC-IV Notes database	Expert consensus	Low	High	Low	Low	Low
2025	Chan[22]	DeepSeek-R1	Not written	Cases from the NEJM	Expert consensus	Low	High	Low	Low	Low
2025	Dai[23]	DeepSeek-R1-Distill-LLaMA-70B	Not written	CTE reports of patients with CD and controls collected in IBD patient registry	Expert consensus	Low	High	Low	Low	Low
2025	Diniz-Freitas[24]	DeepSeek-R1	Not written	The quiz “Image Challenge” in the NEJM	Answer	Low	High	Low	Low	High
2025	Hassanein[25]	DeepSeek-3	Not written	Case reports retrieved in PubMed	Answer	Low	High	Low	Low	High
2025	Ilic[28]	DeepSeek Medical (2024)	Not written	Vignettes obtained from medical records	Expert consensus	Low	High	Low	Low	High
2025	Cheng[29]	DeepSeek-R1	Not written	The case reports from EyeRounds	Expert consensus	Low	Low	Low	Low	High
2025	Kang[30]	DeepSeek-R1	Not written	Cases from the hospital	Expert consensus	Low	High	Low	Low	Low
2025	Kaygisiz[31]	DeepSeek-V3	DeepSeek-2025-02-18	Clinical case scenarios made by the authors	Expert consensus	Unclear	High	Low	Low	High

2025	Kurz[32]	Deepseek VL2 Tiny (1B)	Not written	The quiz “Image Challenge” in the NEJM	Answer	Low	Low	Low	Low	Low
2025	Mikhail[34]	DeepSeek-R1	Not written	Cases from StatPearls	Answer	Low	Low	Low	Low	Low
2025	Moëll[35]	DeepSeek-R1	Not written	Clinical vignettes from the MedQA dataset	Answer	Low	Low	Low	Low	Low
2025	Pan[39]	DeepSeek-R1	Not written	Outpatient cases from hospital	Expert consensus	Low	High	Low	Low	Low
2025	Patel[40]	DeepSeek-R1	Not written	Cases from the Case Book of Sleep Medicine, Third Edition (2019)	Answer	Low	Low	Low	Low	High
2025	Spitzl[41]	DeepSeek-V3	Not written	Fictitious PET/CT reports generated by physicians	Answer	Unclear	High	Low	Low	Low
2025	Tordjman[42]	DeepSeek-R1	Not written	Case challenges in the NEJM	Answer	Low	High	Low	Low	High
2025	Camalan[43]	DeepSeek-V3	Not written	clinical scenarios designed by two otorhinolaryngologists	Answer	Unclear	High	Low	Low	High
2025	Wu[44]	DeepSeek-R1	Not written	Cases published in the NEJM, Mayo Clinic Proceedings, CHEST, Neurology	Answer	Low	Low	Low	Low	High
2025	Li[33]	DeepSeek-V3	Not written	The case reports from EyeRounds	Expert consensus	Low	Low	Low	Low	High



# Artificial Intelligence Applications in Pediatric Genetics: Clinical and Research Implications and Ethical Considerations

 Şerife Öztürk Yılmaz

Department of Pediatric Genetics, Dokuz Eylül University Faculty of Medicine, İzmir, Türkiye

## Abstract

**Introduction:** Artificial intelligence (AI) is transforming healthcare, and pediatric genetics has emerged as a particularly promising field for its implementation due to the complexity of rare genetic disorders and the rapidly increasing volume of genomic data. AI-based technologies may improve diagnostic accuracy, facilitate genomic interpretation, and support precision medicine approaches in pediatric patients.

**Methods:** This study was designed as a structured narrative review informed by a systematic literature search. A comprehensive search was conducted in PubMed, Scopus, and Web of Science databases for English-language publications published between January 2019 and January 2026. Relevant original studies, reviews, consensus statements, and ethically focused publications related to AI applications in pediatric genetics were included.

**Results:** AI-supported tools, including deep learning-based facial dysmorphism analysis and machine learning-driven genomic variant prioritization systems, have demonstrated significant potential in improving phenotype-genotype correlation, diagnostic efficiency, and personalized management strategies. AI applications also contribute to predictive modeling, disease-risk stratification, and large-scale multi-omics research. However, important ethical and legal concerns remain, including informed consent in minors, algorithmic bias, data privacy, transparency, and equitable access to AI technologies.

**Discussion and Conclusion:** AI has considerable potential to transform pediatric genetic practice by enhancing diagnostic precision and supporting individualized clinical care. Nevertheless, successful and equitable integration into routine practice requires robust ethical governance, prospective validation in diverse populations, transparent algorithms, and multidisciplinary clinician oversight. AI should be regarded as a decision-support tool that augments rather than replaces clinical expertise.

**Keywords:** Artificial intelligence; Dysmorphism; Ethical governance; Genomic medicine; Machine learning; Pediatric genetics; Rare diseases; Variant interpretation

Artificial intelligence (AI) has rapidly emerged as a transformative technology in healthcare by enabling advanced data analysis, automated pattern recognition, and predictive modeling.<sup>[1–3]</sup> AI-supported systems

increasingly contribute to diagnostic accuracy, clinical decision-making, and personalized treatment strategies across multiple medical disciplines. Within this landscape, pediatric genetics is a particularly suitable domain for

**Cite this article as:** Öztürk Yılmaz Ş. Artificial Intelligence Applications in Pediatric Genetics: Clinical and Research Implications and Ethical Considerations. Lokman Hekim Health Sci 2026;6(2):334–340.

**Correspondence:** Şerife Öztürk Yılmaz, M.D. Dokuz Eylül Üniversitesi Tıp Fakültesi, Çocuk Genetiği Anabilim Dalı, İzmir, Türkiye

**E-mail:** serifeozturkiyilmaz@gmail.com **Submitted:** 05.03.2026 **Revised:** 18.05.2026 **Accepted:** 11.06.2026 **Available Online:** 16.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



AI implementation, given the complexity of genotype-phenotype relationships and the rapidly expanding volume of genomic data generated by next-generation sequencing technologies.<sup>[4-6]</sup>

Rare genetic disorders collectively affect millions of children worldwide and frequently present with heterogeneous clinical manifestations, overlapping phenotypes, and delayed diagnosis. Despite substantial technological advances in genomic sequencing, many pediatric patients still experience prolonged diagnostic journeys requiring extensive clinical evaluation and repeated testing. AI offers the potential to mitigate these challenges by integrating phenotypic, genomic, and clinical datasets into unified analytical frameworks that support early and accurate diagnosis.<sup>[4,6,7]</sup>

Recent developments in machine learning and deep learning have enabled automated facial dysmorphology recognition, phenotype-driven gene prioritization, and predictive modeling of disease progression.<sup>[1,7-9]</sup> These approaches may significantly improve early diagnosis and individualized management strategies, particularly in childhood-onset genetic disorders where timely intervention can have a profound impact on long-term outcomes.<sup>[1,3,10]</sup>

The primary aim of this review was to provide a clinically oriented synthesis of emerging AI applications in pediatric genetics rather than to quantitatively evaluate pooled evidence.

## Materials and Methods

Eligible publications included original research articles, systematic reviews, and consensus statements with direct clinical applicability to pediatric genetic practice, published in English between January 2019 and January 2026. Foundational studies published before 2019 (e.g., the initial description of the human phenotype ontology (HPO), American College of Medical Genetics and Genomics (ACMG) variant interpretation guidelines, and early phenotype-driven prioritization algorithms) were selectively included when they provided essential conceptual frameworks not superseded by more recent literature. Editorials, conference abstracts, and non-peer-reviewed sources were excluded.

A comprehensive literature search was conducted in January 2026 across PubMed, Scopus, and Web of Science databases. The search strategy combined key terms related to AI ("artificial intelligence," "machine learning," "deep learning") and genetics ("genetics," "genomic medicine," "clinical genetics," "medical genetics") using Boolean

operators. Database-specific adaptations of the search syntax were applied where necessary.

The initial search identified 1,995 records (PubMed: n=301; Web of Science: n=303; Scopus: n=1.391). After removal of duplicates, 1,625 unique records remained. Following title and abstract screening, 1,425 articles were excluded due to lack of relevance to AI applications in genetic or genomic medicine or absence of pediatric applicability. A total of 200 articles were assessed for full-text eligibility. 162 articles were excluded because they primarily focused on non-clinical computational methodologies, lacked relevance to pediatric populations, lacked sufficient methodological detail, or did not address practical applications in genetic or genomic medicine.

This study was designed as a structured narrative review informed by a systematic literature search strategy. The primary objective was to provide a clinically oriented synthesis of current and emerging AI applications in pediatric genetics rather than to perform a quantitative evidence synthesis or meta-analysis.

Because the review included heterogeneous evidence types, including original studies, reviews, consensus statements, methodological reports, and ethical guidance documents, formal systematic review procedures such as preferred reporting items for systematic reviews and meta-analyses methodology, protocol registration, duplicate reviewer screening, and formal risk-of-bias assessment tools were not applied. Instead, the included literature was qualitatively critically appraised for clinical applicability, methodological transparency, population diversity, and relevance to pediatric genetic practice.

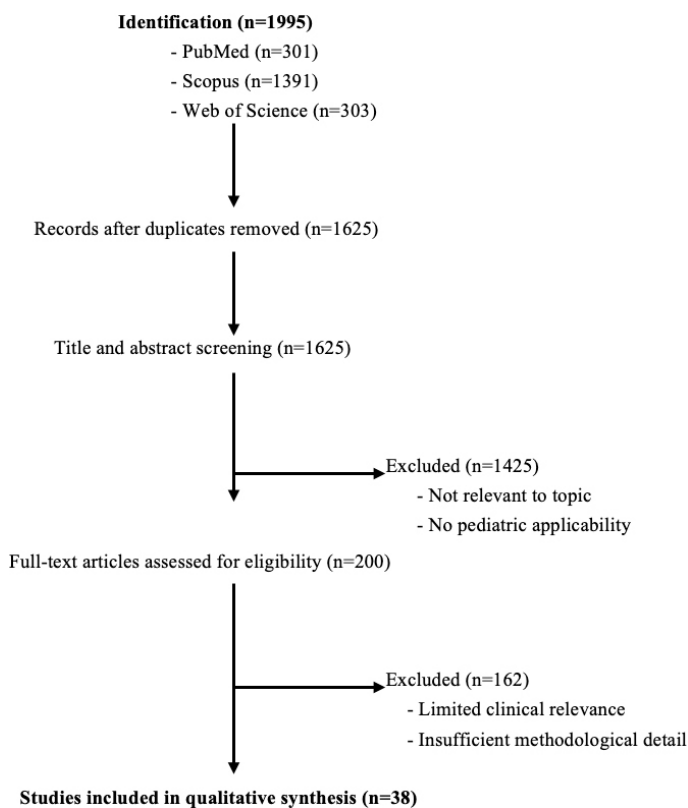
Study selection and eligibility assessment were performed by the sole author. This may represent a methodological limitation and may increase the risk of selection bias.

38 studies were ultimately included in the final review (Fig. 1). Studies not exclusively conducted in pediatric populations were also considered if they demonstrated applications of AI in genetic or genomic analysis with clear translational relevance to pediatric genetic practice. The full list of included studies is provided in Supplementary Table S1.

## Clinical and Research Consequences

A summary of key studies, their AI methods, clinical applications, dataset characteristics, reported performance, and major limitations is presented in Table 1.

The diagnostic evaluation of pediatric rare diseases remains one of the most challenging areas of clinical



**Figure 1.** Preferred reporting items for systematic reviews and meta-analyses 2020 flow diagram of study selection.

medicine. AI-based diagnostic systems enable automated integration of HPO-based clinical data with genomic sequencing results, facilitating phenotype-genotype correlation and reverse phenotyping approaches.<sup>[5,7,9,11]</sup> These tools allow clinicians to reinterpret genomic findings as patients are followed longitudinally, improving diagnostic precision over time.<sup>[2,4,8,12]</sup>

Deep learning-based facial recognition systems have demonstrated high accuracy in identifying syndromic conditions in real-world clinical settings by analyzing dysmorphic facial features.<sup>[5,8,9]</sup> By assisting clinicians in narrowing differential diagnoses and guiding targeted genetic testing, such systems are particularly valuable during early childhood when phenotypic findings may be subtle or non-specific. It is important to emphasize that these tools serve as decision-support systems; even when an AI algorithm suggests a particular syndromic diagnosis, the final clinical determination requires confirmation through comprehensive physical examination and integration with other clinical findings by an experienced clinician.

The widespread clinical adoption of whole-exome and whole-genome sequencing has significantly increased diagnostic yield but also introduced substantial

interpretative complexity. AI-supported variant interpretation platforms integrate multidimensional datasets including evolutionary conservation metrics, functional prediction algorithms, population frequency databases, and phenotype ontologies.<sup>[4,6,13]</sup> Machine learning-based prioritization reduces analytical burden and improves consistency in variant classification in accordance with ACMG recommendations.<sup>[13]</sup> AI-assisted reanalysis further enables iterative diagnostic reassessment as new gene-disease associations are discovered.<sup>[12,14]</sup>

Beyond diagnosis, AI applications are increasingly extending into predictive and preventive pediatric medicine. Integrative AI models combining genomic, biochemical, imaging, and longitudinal clinical data allow prediction of disease progression and individualized risk assessment.<sup>[1-3]</sup> Such predictive systems may support early identification of clinical deterioration, optimize surveillance strategies, and enable personalized therapeutic decision-making in inherited metabolic, neurodevelopmental, and cancer predisposition disorders. Taken together, these developments contribute to a continuously learning healthcare system in which genomic data remain clinically informative throughout childhood and adulthood.<sup>[2,10,15]</sup>

AI technologies are reshaping genetic research by enabling large-scale multi-omics analysis and discovery of novel disease mechanisms.<sup>[10,15,16]</sup> Machine learning approaches facilitate the identification of gene-disease associations and biological pathways that may remain undetected using conventional analytical methods. Collaborative frameworks such as federated learning allow model development across multiple institutions without direct sharing of sensitive genomic data, improving generalizability while preserving patient privacy and regulatory compliance.<sup>[17,18]</sup> This approach has the additional advantage of enabling algorithm training on more diverse datasets, potentially reducing algorithmic bias and improving performance across different populations.

## Ethical and Legal Challenges

The clinical implementation of AI in pediatric genetics introduces significant ethical and legal challenges. Genomic information represents permanently identifiable health data with implications extending across an individual's lifespan and potentially affecting family members.<sup>[4,12,17]</sup> Pediatric genomic testing further raises concerns because consent decisions are typically made by parents or guardians while long-term consequences affect the child's future autonomy, commonly described as the

**Table 1.** Key applications, foundational studies, and methodological considerations of artificial intelligence in pediatric genetic practice

Study (Year)	AI method	Clinical application	Dataset characteristics	Key reported performance	Major limitation(s)
Gurovich et al. <sup>[8]</sup>	Deep learning (CNN)	Facial dysmorphology analysis for syndromic diagnosis	~17,000 images; predominantly European ancestry	Top-10 accuracy: 91%	Limited ethnic diversity; single-modality (facial images only)
Hsieh et al. <sup>[9]</sup>	Deep learning (Gestalt Matcher)	Rare disease matching using facial phenotype descriptors	~30,000 images; multi-ethnic but imbalanced	AUC: 0.91 for known syndromes	Performance drops significantly for ultra-rare syndromes; requires further prospective validation
Schaefer et al. <sup>[11]</sup>	Various ML methods	Scoping review of ML in rare diseases	211 included studies	N/A (review)	Heterogeneous methodologies; limited clinical implementation data
Dias and Torkamani <sup>[6]</sup>	Machine learning	Genomic variant prioritization	Multiple datasets (simulated and real)	Performance varies across models; reported improvements in variant prioritization accuracy compared to traditional pipelines	Lack of standardized benchmarking and persistent “black box” interpretability challenges
Richards et al. <sup>[13]</sup>	N/A (guidelines)	ACMG variant interpretation guidelines	N/A (consensus)	N/A	Pre-AI framework; does not incorporate AI-based evidence or probabilistic modeling approaches
Teo et al. <sup>[18]</sup>	Federated learning	Systematic review of federated ML in healthcare	612 studies	N/A (review)	Limited real-world clinical implementation; interoperability and data privacy constraints
Obermeyer et al. <sup>[19]</sup>	Algorithmic bias analysis	Racial bias in healthcare algorithm	~50,000 patients	Significant racial bias detected	Not specific to genomics or pediatric populations; highlights broader risks of algorithmic bias in clinical AI systems

AI: Artificial intelligence; AUC: Area under the curve; CNN: Convolutional neural network; ML: Machine learning; N/A: Not applicable.

child's “right to an open future.”<sup>[4,12,16]</sup> This concept creates particular tension in the context of AI-based predictive analytics: when algorithms identify risks for adult-onset conditions during childhood, clinicians and families must navigate the complex decision of whether and how to disclose information that may impact the child's future autonomous decision-making.

## Management of Incidental and Secondary Findings

AI-driven genomic analysis, particularly when using broad variant-prioritization algorithms, may inadvertently uncover incidental or secondary findings unrelated to the initial diagnostic question.<sup>[12,16]</sup> These may include adult-onset conditions, carrier status for recessive disorders, or even misattributed parentage. Clinicians using AI tools must therefore establish clear protocols for handling incidental findings, including predefined

policies on when, how, and whether to disclose this information to families, balancing immediate clinical utility with the child's future autonomy.<sup>[4]</sup>

## Data Ownership and Long-Term Data Stewardship

Pediatric genomic data used to train or validate AI models cannot be fully anonymized due to their inherent uniqueness.<sup>[17]</sup> This raises critical questions regarding data ownership: Who holds the rights to a child's genomic data after it is used to train a commercial AI algorithm? Current regulatory frameworks, including general data protection regulation, provide a “right to be forgotten,” but its technical implementation for already-trained AI models remains unresolved.<sup>[16]</sup> Commercial AI developers may have conflicting incentives regarding long-term data stewardship, creating potential tensions between profit motives and ethical obligations to research participants.

## Obligations for Longitudinal Reanalysis

AI algorithms improve over time as new data and methodological advances emerge. Studies have demonstrated that AI-assisted reanalysis can identify new diagnoses in a substantial proportion of previously unsolved cases.<sup>[12,14]</sup> However, without clear guidelines, a two-tiered system may emerge where patients at academic medical centers benefit from continuous reanalysis while those in resource-limited settings do not, potentially exacerbating existing healthcare disparities. Establishing systematic reanalysis pipelines requires significant infrastructure, funding, and personnel – an obligation that remains largely unaddressed in current policy frameworks.

## Algorithmic Bias and Transparency

Algorithmic bias remains a major concern, as AI systems trained on genetically homogeneous datasets may perform worse in underrepresented populations and unintentionally reinforce healthcare disparities.<sup>[17,19]</sup> Furthermore, limited transparency of complex AI models – the so-called “black box” problem – complicates attribution of medicolegal responsibility among clinicians, healthcare institutions, and AI developers, highlighting the importance of explainable AI in clinical decision-making.<sup>[3,17,20]</sup>

## Limitations and Methodological Considerations

The findings and claims discussed above must be interpreted in light of several important limitations. First, while deep learning-based facial recognition systems have demonstrated high diagnostic accuracy in controlled research settings,<sup>[5,8,9]</sup> it is critical to note that most training datasets are predominantly composed of individuals of European ancestry. For example, Gurovich and colleagues reported 91% top-10 accuracy using facial analysis<sup>[8]</sup> yet their model’s generalizability to more diverse populations – including individuals of African, Asian, or Latin American descent – remains unproven, a limitation explicitly acknowledged by the original authors. This raises the possibility that reported accuracy figures may overestimate real-world performance in non-European populations.

Second, AI-supported variant interpretation platforms rely heavily on reference databases (e.g., gnomAD, ClinVar) that exhibit substantial ancestry bias, with underrepresented populations having higher rates of variants of uncertain significance. Consequently, diagnostic yield improvements reported in the literature<sup>[4,6,13]</sup> may not be equitably

distributed across all pediatric patient populations. Third, the “black box” problem mentioned earlier complicates not only medicolegal attribution but also clinical adoption in resource-limited settings where explainability is particularly valued.

Fourth, most published studies on AI in pediatric genetics focus on retrospective cohorts or curated datasets, which may not reflect the phenotypic heterogeneity and data incompleteness encountered in routine clinical practice. Prospective validation studies in unselected, consecutive patient populations remain scarce. Fifth, the promising predictive analytics applications described above<sup>[1–3]</sup> are largely derived from single-center or simulation-based studies; their performance in real-time clinical decision support, particularly for progressive or ultra-rare conditions, requires further prospective evaluation.

Finally, our discussion is limited by the absence of a systematic review methodology, and the cited references serve as illustrative examples rather than an exhaustive synthesis of the literature. Future research should prioritize diverse, multi-ethnic training cohorts, standardized reporting of subgroup performance metrics, and prospective clinical trials to establish generalizable evidence for AI-based diagnostic tools in pediatric genetics. In addition, publication bias may have influenced the available evidence, as studies reporting positive performance outcomes are more likely to be published. Study selection was conducted by a single reviewer, which may increase the risk of selection bias. The evidence discussed in this review includes heterogeneous publication types with varying methodological strengths; therefore, conclusions should be interpreted within the context of the underlying evidence level.

Despite these challenges, AI-supported tools have the potential to substantially improve daily clinical practice in pediatric genetics by facilitating earlier diagnosis, optimizing genomic data interpretation, and supporting individualized patient management strategies. AI-assisted phenotype recognition and variant prioritization may reduce diagnostic delays and enable more efficient utilization of genetic testing resources. Predictive analytics may further assist clinicians in risk stratification, longitudinal monitoring, and personalized surveillance planning for children with genetic disorders.<sup>[1–3]</sup> Successful implementation ultimately requires clinician oversight, interdisciplinary collaboration, and integration into existing clinical workflows to ensure safe, equitable, and patient-centered care.

## Conclusion

AI has the potential to significantly transform pediatric genetic practice, improving diagnostic efficiency, enhancing genomic interpretation, and enabling precision-based patient management. However, the successful integration of AI into clinical workflows depends on careful attention to ethical governance, data security, algorithmic fairness, and transparency. Rather than replacing clinical expertise, AI should be regarded as a decision-support partner that augments diagnostic reasoning and clinical judgment.

Future research should specifically: (1) Conduct prospective, multi-center validation of at least two existing facial recognition tools (e.g., GestaltMatcher and DeepGestalt) in cohorts where <30% of participants are of European ancestry; (2) report performance metrics stratified by ancestry, age, and phenotypic complexity in all AI studies; (3) develop and publicly release a standardized benchmarking dataset for pediatric genomic AI that includes diverse populations; and (4) establish within the next 5 years a regulatory framework requiring prospective clinical trial data for AI-based diagnostic tools before regulatory approval in pediatric populations.

Only through rigorous evaluation and responsible governance can the full potential of AI be safely and equitably realized for children with genetic disorders and their families.

**Ethics Committee Approval:** Ethical approval was not required for this study since this is a review article.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** None declared.

**Peer-review:** Double blind peer-reviewed.

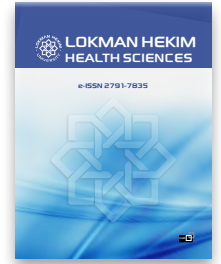
## References

- Schaefer J, Lehne M, Schepers J, Prasser F, Thun S. The use of machine learning in rare diseases: a scoping review. *Orphanet J Rare Dis* 2020;15:145. [\[CrossRef\]](#)
- Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380(14):1347-58. [\[CrossRef\]](#)
- Marques M, Almeida A, Pereira H. The medicine revolution through artificial intelligence: ethical challenges of machine learning algorithms in decision-making. *Cureus* 2024;16(9):e69405. [\[CrossRef\]](#)
- Ilić N, Sarajlija A. Artificial intelligence in the diagnosis of pediatric rare diseases: from real-world data toward a personalized medicine approach. *J Pers Med* 2025;15(9):407. [\[CrossRef\]](#)
- Duong D, Solomon BD. Artificial intelligence in clinical genetics. *Eur J Hum Genet.* 2025;33(3):281-8. [\[CrossRef\]](#)
- Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med.* 2019;11:70. [\[CrossRef\]](#)
- Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med* 2015;7(1):81. [\[CrossRef\]](#)
- Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med* 2019;25(1):60-4. [\[CrossRef\]](#)
- Hsieh TC, Bar-Haim A, Moosa S, Ehmke N, Gripp KW, Pantel JT, et al. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nat Genet* 2022;54(3):349-57. [\[CrossRef\]](#)
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44-56. [\[CrossRef\]](#)
- Robinson PN, Mundlos S. The human phenotype ontology. *Clin Genet* 2010;77(6):525-34. [\[CrossRef\]](#)
- Krier JB, Kalia SS, Green RC. Genomic sequencing in clinical practice: applications, challenges, and opportunities. *Dialogues Clin Neurosci* 2016;18(3):299-312. [\[CrossRef\]](#)
- Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17(5):405-24. [\[CrossRef\]](#)
- Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319(13):1317-8. [\[CrossRef\]](#)
- Emmert-Streib F, Bottini S, Franco L. Editorial: AI and multi-omics for rare diseases: challenges, advances and perspectives, Volume III. *Front Mol Biosci* 2024;11:1392943. [\[CrossRef\]](#)
- World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO; 2024.
- He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25(1):30-6. [\[CrossRef\]](#)
- Teo ZL, Jin L, Li S, et al. Federated machine learning in healthcare: a systematic review on clinical applications and technical architecture. *Cell Rep Med* 2024;5(2):101419. [\[CrossRef\]](#)
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366(6464):447-53. [\[CrossRef\]](#)
- Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020;20(1):310. [\[CrossRef\]](#)
- Amiri H, Kohane IS. Machine Learning of Patient Characteristics to Predict Admission Outcomes in the Undiagnosed Diseases Network. *JAMA Netw Open* 2021;4(2):e2036220. [\[CrossRef\]](#)
- Arlt A, Knaus A, Hsieh TC, Klinkhammer H, Bhasin MA, Hustinx A, et al. Next-generation phenotyping in Nigerian children with Cornelia de Lange syndrome. *Am J Med Genet A* 2024;194(9):e63641. [\[CrossRef\]](#)

23. Berger SI, Pitsava G, Cohen AJ, Délot EC, LoTempio J, Andrew EH, et al. Increased diagnostic yield from negative whole genome-slice panels using automated reanalysis. *Clin Genet*. 2023;104:377-83. [\[CrossRef\]](#)
24. Beyreli I, Karakahya O, Cicek AE. DeepND: Deep multitask learning of gene risk for comorbid neurodevelopmental disorders. *Patterns (N Y)* 2022;3(7):100524. [\[CrossRef\]](#)
25. Jackson S, Freeman R, Noronha A, Jamil H, Chavez E, Carmichael J, et al. Applying data science methodologies with artificial intelligence variant reinterpretation to map and estimate genetic disorder prevalence utilizing clinical data. *Am J Med Genet A* 2024;194(5):e63505. [\[CrossRef\]](#)
26. Kabir M, Stuart HM, Lopes FM, Fotiou E, Keavney B, Doig AJ, et al. Predicting congenital renal tract malformation genes using machine learning. *Sci Rep* 2023;13:13204. [\[CrossRef\]](#)
27. Kaczmarek AT, Bahlmann N, Thaqi B, May P, Schwarz G. Machine learning-based identification and characterization of 15 novel pathogenic SUOX missense mutations. *Mol Genet Metab* 2021;134(1-2):188-94. [\[CrossRef\]](#)
28. Kadlubowska MK, Schrauwen I. Methods to Improve Molecular Diagnosis in Genomic Cold Cases in Pediatric Neurology. *Genes (Basel)* 2022;13(2):333. [\[CrossRef\]](#)
29. Kingsmore SF, Wright M, Smith LD, Liang Y, Mowrey WR, Protopsaltis L, et al. Prequalification of genome-based newborn screening for severe childhood genetic diseases through federated training based on purifying hyperselection. *Am J Hum Genet* 2024;111(12):2618-42. [\[CrossRef\]](#)
30. Kobayashi Y, Chen E, Facio FM, Metz H, Poll SR, Swartzlander D, et al. Clinical variant reclassification in hereditary disease genetic testing. *JAMA Netw Open* 2024;7(11):e2444526. [\[CrossRef\]](#)
31. Peterson B, Hernandez EJ, Hobbs C, Malone Jenkins S, Moore B, Rosales E, et al. Automated prioritization of sick newborns for whole genome sequencing using clinical natural language processing and machine learning. *Genome Med* 2023;15(1):18. [\[CrossRef\]](#)
32. Porras AR, Rosenbaum K, Tor-Diez C, Summar M, Linguraru MG. Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: a multinational retrospective study. *Lancet Digit Health* 2021;3(10):e635-43. [\[CrossRef\]](#)
33. Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, et al. Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes. *Genet Med* 2022;24(6):1336-48. [\[CrossRef\]](#)
34. Rivera-Munoz EA, Zhao XE, Rosenfeld JA, Luna PN, Shaw CA, Posey JE, et al. Clinical exome sequencing efficacy and phenotypic expansions involving non-isolated congenital anomalies of kidney and urinary tract (CAKUT+). *Eur J Hum Genet* 2025;33(12):1606-15. [\[CrossRef\]](#)
35. Yang Z, Shikany A, Ni Y, Zhang G, Weaver KN, Chen J. Using deep learning and electronic health records to detect Noonan syndrome in pediatric patients. *Genet Med* 2022;24(11):2329-37. [\[CrossRef\]](#)
36. Ye GX, Ontiveros E, Ivander A, Velinov M, Simotas C. Autosomal Recessive Infantile Hyaline Fibromatosis Identified Using Artificial Intelligence-Assisted Rapid Whole Genome Sequencing: A Rare, Multisystemic, Hereditary Disorder. *Cureus* 2024;16(6):e62037. [\[CrossRef\]](#)
37. Coghlan S, Gyngell C, Vears DF. Ethics of artificial intelligence in prenatal and pediatric genomic medicine. *J Community Genet* 2024;15(1):13-24. [\[CrossRef\]](#)
38. Gripp KW. Artificial intelligence supported facial feature analysis in medical genetics. *Curr Opin Pediatr* 2025;37(6):533-7. [\[CrossRef\]](#)

<b>Appendix 1. Summary of included studies and key supporting literature on artificial intelligence in pediatric genomic medicine</b>				
<b>No</b>	<b>Study</b>	<b>Type</b>	<b>AI method</b>	<b>Clinical application/relevance</b>
1	Gurovich et al. <sup>[8]</sup>	Original study	Deep learning	Facial dysmorphology analysis
2	Hsieh et al. <sup>[9]</sup>	Original study	Deep learning	Rare disease matching via facial phenotype
3	Amiri H and Kohane <sup>[21]</sup>	Original study	Machine learning	Outcome prediction in undiagnosed diseases
4	Arlt et al. <sup>[22]</sup>	Original study	Deep learning	Facial phenotyping in CdLS
5	Berger et al. <sup>[23]</sup>	Original study	Automated reanalysis	Variant reinterpretation
6	Beyreli et al. <sup>[24]</sup>	Original study	Deep multitask learning	Gene risk prediction (NDD)
7	Jackson et al. <sup>[25]</sup>	Original study	AI-based variant reinterpretation	Genetic disease prevalence estimation
8	Kabir et al. <sup>[26]</sup>	Original study	Machine learning	Congenital anomaly gene prediction
9	Kaczmarek et al. <sup>[27]</sup>	Original study	Machine learning	Pathogenic variant identification
10	Kadlubowska and Schrauwen <sup>[28]</sup>	Review	—	Molecular diagnosis in pediatric neurology
11	Kingsmore et al. <sup>[29]</sup>	Original study	Federated learning	Newborn genomic screening
12	Kobayashi et al. <sup>[30]</sup>	Original study	AI-based analysis	Clinical variant reclassification
13	Peterson et al. <sup>[31]</sup>	Original study	NLP+ML	Prioritization for WGS
14	Porras et al. <sup>[32]</sup>	Original study	Machine learning	Genetic syndrome screening
15	Cohen et al. <sup>[33]</sup>	Original study	Genomic analytics	Rare disease genome analysis
16	Rivera-Munoz et al. <sup>[34]</sup>	Original study	Exome sequencing	Phenotypic expansion
17	Yang et al. <sup>[35]</sup>	Original study	Deep learning	Noonan syndrome detection
18	Ye et al. <sup>[36]</sup>	Original study	AI+rapid WGS	Rare disease diagnosis
19	Obermeyer et al. <sup>[19]</sup>	Analytical study	Algorithmic analysis	Bias in healthcare AI
20	Schaefer et al. <sup>[1]</sup>	Review	—	Machine learning in rare diseases
21	Rajkomar et al. <sup>[2]</sup>	Review	—	Machine learning in medicine
22	Marques et al. <sup>[3]</sup>	Review	—	Ethical challenges of AI
23	Ilić and Sarajlija <sup>[4]</sup>	Review	—	AI in pediatric rare disease diagnosis
24	Duong and Solomon <sup>[5]</sup>	Review	—	AI in clinical genetics
25	Dias and Torkamani <sup>[6]</sup>	Review	—	AI in genomic diagnostics
26	Smedley and Robinson <sup>[7]</sup>	Methodological	—	Phenotype-driven gene prioritization
27	Topol <sup>[10]</sup>	Perspective	—	AI in medicine framework
28	Robinson and Mundlos <sup>[11]</sup>	Methodological	—	Human Phenotype Ontology
29	Krier et al. <sup>[12]</sup>	Review	—	Genomic sequencing in clinical practice
30	Richards et al. <sup>[13]</sup>	Guideline	—	ACMG variant interpretation standards
31	Beam and Kohane <sup>[14]</sup>	Perspective	—	Big data and AI in healthcare
32	Emmert-Streib et al. <sup>[15]</sup>	Editorial	—	AI and multi-omics
33	World Health Organization <sup>[16]</sup>	Guideline	—	Ethics and governance of AI
34	He et al. <sup>[17]</sup>	Review	—	AI implementation in medicine
35	Teo et al. <sup>[18]</sup>	Systematic review	—	Federated learning in healthcare
36	Amann et al. <sup>[20]</sup>	Review	—	Explainability in AI
37	Coghlan et al. <sup>[37]</sup>	Review	—	Ethics in pediatric genomic AI
38	Gripp <sup>[38]</sup>	Review	—	AI facial analysis in genetics

This table includes heterogeneous evidence types, including original studies, reviews, guidelines, methodological reports, perspectives, and editorials, to provide a clinically oriented overview of current AI applications in pediatric genomic medicine. ACMG: American College of Medical Genetics and Genomics; AI: Artificial intelligence; CdLS: Cornelia de Lange syndrome; WGS: Whole genome sequencing; NDD: Neurodevelopmental disorders.



# Artificial Intelligence as a Managerial Decision Support Infrastructure in Hospitals: A Governance Framework

İsmet Şahin<sup>1</sup>, Ferhat Devrim Zengul<sup>2</sup>

<sup>1</sup>Department of Health Services Management, Lokman Hekim University, Ankara, Türkiye

<sup>2</sup>Department of Health Services Administration, Faculty of Health Professions, University of Alabama at Birmingham, Birmingham, AL, United States

## Abstract

The advent of artificial intelligence (AI) has precipitated a paradigm shift in healthcare; however, scholarly and policy discourse still prioritizes clinical applications, while managerial applications remain comparatively underexamined. This narrative review focuses on AI governance in hospital management and repositions AI as a managerial decision support infrastructure capable of reshaping hospital governance, organizational accountability, and institutional legitimacy. We conducted a structured narrative synthesis drawing on organizational theory, healthcare management, and AI governance literature. Relevant sources were identified through targeted database searches and citation tracking, screened for applicability to hospital-level managerial AI (strategic planning, resource allocation, and performance oversight), and synthesized via iterative thematic analysis to identify recurring governance challenges and convergent mechanisms. Based on this synthesis, we develop a three-dimensional governance framework linking managerial AI to (1) decision-authority distribution, (2) accountability mechanisms, and (3) institutional legitimacy. The framework is operationalized with illustrative scenarios and contextualized for the Turkish healthcare system. Our analysis shows that effective managerial AI governance requires explicit authority assignment, baseline thresholds for transparency and auditability, and board-level oversight structures to prevent responsibility diffusion and protect institutional legitimacy. We conclude with implementation guidance for hospital leaders and health system regulators.

**Keywords:** Artificial intelligence governance; Artificial intelligence; Healthcare Management; Hospital governance; Managerial decision support; Narrative review

Artificial intelligence (AI) has become integral to healthcare systems, with substantial literature on the diagnostic accuracy and patient safety of its clinical applications.<sup>[1–4]</sup> This clinically oriented scholarship has shaped ethical standards and regulatory frameworks for AI in direct patient care.

Hospitals depend on strategic, financial, and operational decisions regarding resource allocation, workforce planning, and service configuration that fundamentally shape care delivery.<sup>[5,6]</sup> Increasingly, these organizational decisions are informed by AI systems that serve as managerial decision-support infrastructure, influencing

**Cite this article as:** Şahin İ, Zengul FD. Artificial Intelligence as a Managerial Decision Support Infrastructure in Hospitals: A Governance Framework. Lokman Hekim Health Sci 2026;6(2):341–351.

**Correspondence:** İsmet Şahin, M.D. Lokman Hekim Üniversitesi, Sağlık Hizmetleri Yönetimi Bölümü, Ankara, Türkiye

**E-mail:** ismet.sahin@lokmanhekim.edu.tr **Submitted:** 21.01.2026 **Revised:** 08.02.2026 **Accepted:** 17.02.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



executive and board-level choices rather than merely assisting clinicians at the point of care.

Despite this shift, the governance implications of managerial AI remain insufficiently theorized. Managerial decisions differ fundamentally from clinical decisions in scope, stakeholder complexity, and accountability expectations. While clinical AI operates within established professional governance frameworks (medical liability, clinical standards, professional ethics), managerial AI influences organizational priorities and system-level equity without precise accountability mechanisms.<sup>[7–9]</sup> This creates what organizational scholars describe as an accountability gap diffusion of responsibility across human decision-makers, algorithmic systems, and data infrastructures that challenge traditional governance arrangements.<sup>[10,11]</sup>

The governance vacuum is particularly consequential in healthcare organizations with strong public-interest obligations. When AI systems inform decisions about service availability, staffing levels, or resource distribution, the downstream effects extend to patient access, health equity, and institutional legitimacy. Yet existing governance frameworks – designed for clinical AI or adapted from corporate governance models – do not fully address these organizational and societal accountability challenges.

We define the accountability gap as the diffusion of responsibility across decision-makers, model builders, and data infrastructures, such that no single actor can provide complete justification or be held accountable for downstream harms arising from AI-supported organizational decisions. This operational definition guides the development of our framework.

Research question: How should hospitals govern AI systems that function as managerial decision support infrastructure, and what accountability mechanisms prevent responsibility diffusion while maintaining institutional legitimacy?

We: (1) distinguish managerial AI from clinical AI across key organizational dimensions; (2) examine how managerial AI reconfigures decision authority and creates accountability gaps; (3) propose a governance framework linking managerial AI to accountability mechanisms and institutional legitimacy; and (4) contextualize implications for hospital governance in the Turkish healthcare system. By repositioning managerial AI usage as a governance challenge rather than a technical challenge, this review contributes theoretical guidance for health sciences scholarship and practical frameworks for hospital leaders.

## Materials and Methods

### Review Design and Scope

In this study, we followed narrative review methodology<sup>[12,13]</sup> and prioritized theoretical integration and conceptual synthesis over systematic evidence aggregation. This narrative review develops a governance framework for managerial AI using concept-driven synthesis appropriate for emergent, heterogeneous topics<sup>[12,13]</sup> and synthesizes literature to develop a governance framework for managerial AI in hospitals. The review addresses a multidisciplinary problem that requires interpretive analysis across organizational theory, healthcare management, and AI governance.

### Literature Identification and Selection

Sources were identified through structured searches in PubMed, Web of Science, Scopus, and Google Scholar (January 2015–December 2025). Search terms combined: (1) AI terms: AI, machine learning, algorithmic decision-making; (2) organizational terms: Hospital management, healthcare governance, organizational decision-making; and (3) governance terms: accountability, decision authority, institutional legitimacy. Seminal theoretical works (e.g., Simon, Mintzberg) were included through citation tracking. Inclusion: peer-reviewed articles on AI in organizational/managerial contexts, healthcare governance, or algorithmic accountability. Exclusion: purely clinical AI or non-English sources.

Iterative selection involved reviewing approximately 150 full-text articles identified through database searches and citation tracking. Selection criteria operationalized “theoretical depth” as: (a) articles proposing or testing organizational theories relevant to algorithmic decision-making; (b) empirical studies examining accountability or governance in technology-intensive organizations; or (c) conceptual analyses addressing authority distribution in human-AI systems. Final inclusion of sources prioritized recent empirical evidence (2020–2024), foundational theoretical contributions from organizational and accountability scholarship, and governance-oriented analyses directly relevant to healthcare settings. To minimize selection bias, a conscious decision was made to include a range of sources that represented divergent perspectives. This approach encompassed the incorporation of optimistic views on algorithmic objectivity, alongside critical perspectives on bias. In addition, arguments for proprietary protection were balanced with calls for transparency, and technological positions were juxtaposed with institutional governance emphasis. The deliberate inclusion of contradictory

**Table 1.** Organizational distinctions between clinical AI and managerial AI in hospital settings

Dimension	Clinical AI	Managerial AI
Primary purpose	Diagnosis, treatment optimization, and patient outcome prediction	Strategic planning, resource allocation, and performance monitoring
Decision level	Individual patient	Organizational and system level
Primary users	Physicians, nurses, allied health professionals	Hospital executives, senior managers, governing boards
Data sources	Clinical records, imaging, laboratory results, physiological signals	Administrative, financial, operational, workforce, performance data
Accountability framework	Clinical governance, professional standards, medical liability	Organizational governance, board oversight, public accountability
Risk profile	Patient safety risks, diagnostic errors, treatment bias	Accountability gaps, erosion of legitimacy, systemic inequities
Transparency focus	Clinical explainability for individual decisions	Decision authority traceability, governance processes
Ethical emphasis	Patient autonomy, beneficence, non-maleficence	Fairness, accountability, legitimacy, public trust

This table synthesizes distinctions identified across organizational theory,<sup>[26,43]</sup> healthcare management,<sup>[5,6,33]</sup> and AI governance literature.<sup>[33,37,38]</sup> The dimensions were selected based on their governance relevance, rather than on technical characteristics of AI systems. AI: Artificial intelligence.

viewpoints ensures that the framework reflects contested terrain rather than privileging a single perspective.

### Analysis Approach

Literature was analyzed thematically<sup>[14]</sup> to identify: (1) distinguishing features of managerial versus clinical AI; (2) governance and accountability challenges; and (3) existing governance frameworks and their limitations.<sup>[15-17]</sup> Synthesis focused on conceptual integration to ensure the proposed framework is theoretically grounded in organizational theory and responsive to hospital management realities. The Turkish healthcare context was incorporated through policy documents and comparative health systems analysis.

## Managerial AI as Decision Support Infrastructure: A Conceptual Framework

### Defining Managerial AI

Managerial AI refers to algorithmic systems supporting organizational decision-making. AI encompasses machine learning and predictive analytics that process large-scale data, identify patterns, and inform strategic decisions that affect multiple stakeholders in hospital settings.<sup>[18,19]</sup>

Managerial AI has indirect effects on patient care, positioning it at the efficiency-equity intersection with distinct governance challenges.

### Distinguishing Managerial and Clinical AI

Table 1 systematically contrasts clinical AI and managerial AI across key organizational dimensions, highlighting why governance frameworks designed for clinical AI do not

translate to managerial contexts. Clinical AI is embedded in established professional accountability mechanisms (e.g., medical liability and clinical governance) and primarily informs decisions affecting individual patient trajectories within ethical frameworks emphasizing autonomy, beneficence, and non-maleficence. In contrast, managerial AI influences strategic planning, resource allocation, and performance oversight – domains where accountability is organizationally distributed, oversight structures are often less explicit, and decisions shape institutional priorities with system-level equity and legitimacy implications. These differences underscore the need for governance arrangements that prioritize decision-authority traceability, institutional oversight, and public accountability in AI-supported managerial decision-making.

### Illustrative Scenarios: Managerial AI in Practice

The following scenarios are analytically constructed, illustrative cases designed to synthesize governance challenges reported across healthcare management and AI governance literature, as well as recurring themes observed in documented organizational incidents. They are not presented as detailed empirical case studies of specific hospitals. Instead, they abstract commonly reported patterns in AI-supported managerial decision-making to concretize the conceptual tensions discussed in this article – particularly accountability gaps, authority ambiguities, and transparency failures. Their purpose is therefore not empirical validation, but to clarify how governance breakdowns can emerge in everyday managerial AI use.

Scenario 1: Emergency department (ED) capacity planning and equity implications.

A university hospital deploys a machine-learning model to support ED staffing and bed-flow planning. Over several months, the model repeatedly underestimates weekend demand patterns in catchment areas with higher social deprivation, contributing to longer waiting times and crowding during predictable peaks. When concerns are raised, operational leaders argue that the model's outputs were "data-driven" and should be followed to maintain efficiency, while IT teams claim they only implemented what was requested.

Governance question: Who is accountable when AI-supported managerial decisions generate systematic inequities, and what oversight mechanisms ensure that efficiency optimization does not undermine equity obligations?

Scenario 2: Workforce optimization, board approval, and patient safety spillovers.

A managerial AI tool recommends reducing night staffing in a geriatric ward by 15% based on historical utilization and cost indicators. The recommendation is escalated as a strategic efficiency measure and approved at the governance level, with the rationale that the system improves resource allocation. Within subsequent months, patients fall, and adverse events rise substantially, prompting internal disputes. Executives highlight that the board approved the decision; board members argue that management proposed the change; analysts and data scientists emphasize that the system only provided recommendations.

Governance question: How should decision authority be distributed and documented when AI-informed strategic choices have direct patient safety implications, and what decision protocols ensure that algorithmic recommendations do not displace accountable human judgment?

Scenario 3: Performance dashboards, proprietary models, and institutional legitimacy.

A teaching hospital introduces an AI-enabled performance dashboard producing departmental "efficiency scores" to guide budget negotiations and managerial evaluations. The scoring method relies on proprietary vendor algorithms that are not fully disclosed, and when departments contest their scores, leadership cannot provide a comprehensible explanation of how outputs were derived. This opacity generates distrust, resistance, and claims that governance is being replaced by vendor-defined criteria rather than institutional values.

Governance question: What transparency and explainability standards are necessary to preserve institutional legitimacy

when proprietary AI systems influence performance assessment and resource decisions?

Taken together, these illustrative scenarios demonstrate that accountability gaps in managerial AI emerge not solely from technical limitations but from institutional design failures, including responsibility diffusion, unclear authority structures, and limited transparency. They underscore the need for governance arrangements that make explicit (i) who owns decisions, (ii) how algorithmic inputs are reviewed and overridden, and (iii) what baseline thresholds for transparency are required for meaningful oversight and stakeholder trust.

## Governance Challenges in Managerial AI Adoption

### The Accountability Gap

Organizational accountability requires identifiable decision-makers who justify choices and accept responsibility.<sup>[20,21]</sup> Managerial AI distributes influence across executives, data scientists, and algorithms – none of whom bear full responsibility.<sup>[10,11,22]</sup>

This creates responsibility gaps where harm occurs without clear accountability,<sup>[23]</sup> undermining institutional legitimacy when AI produces inequitable outcomes.

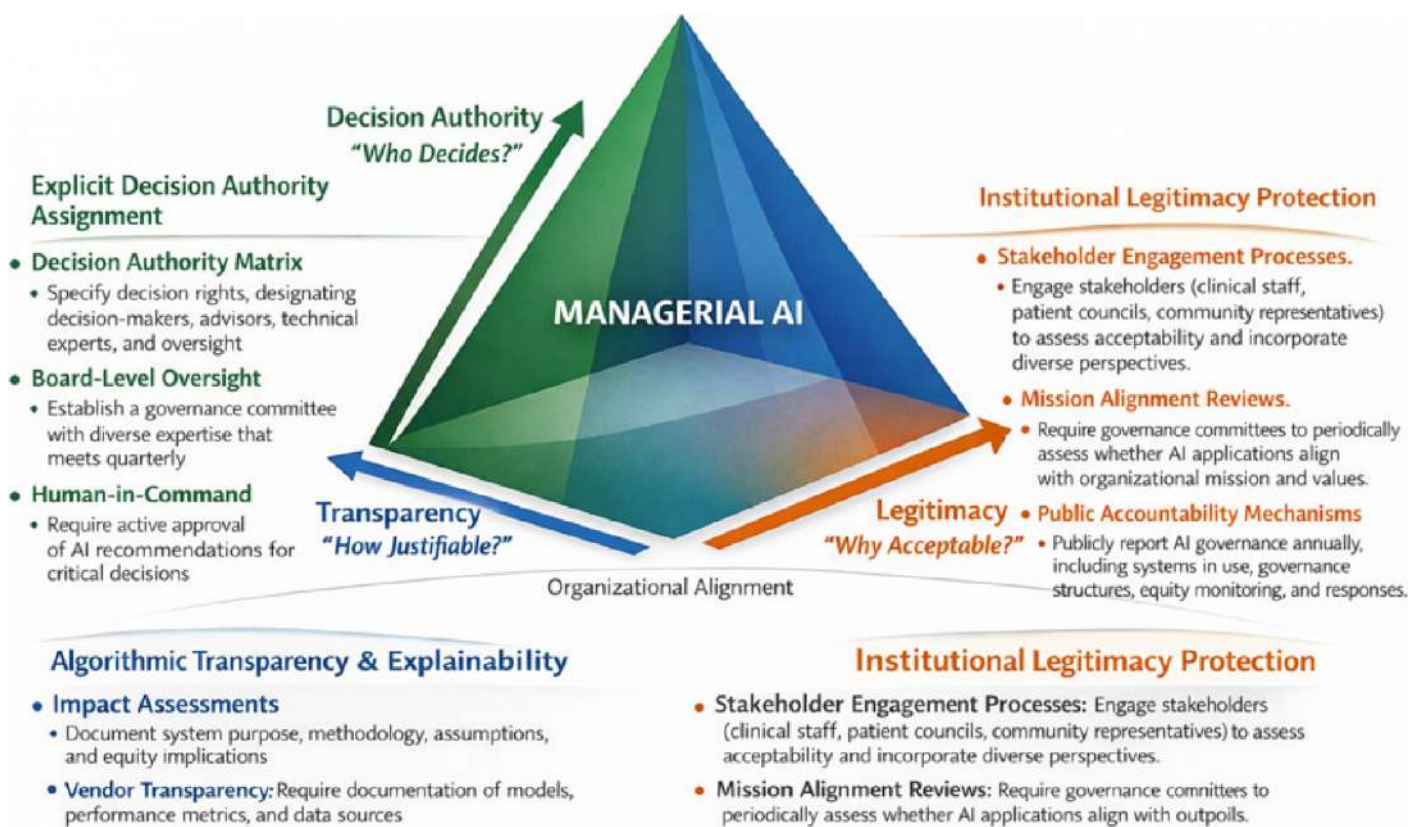
### Opacity and the Black Box Problem

Many managerial AI systems operate as "black boxes" with opaque processes.<sup>[24,25]</sup> This poses three governance challenges:

First, decision-makers cannot adequately evaluate AI recommendations if they do not understand how conclusions were reached. Boards and executives asked to approve strategic decisions based on algorithmic analysis may accept recommendations uncritically, abdicating judgment to systems they cannot interrogate.<sup>[26,27]</sup>

Second, opacity prevents stakeholders from identifying embedded biases or questionable assumptions. Algorithms trained in historical data may perpetuate existing inequities – such as systematically underserving marginalized populations – without detection.<sup>[28,29]</sup> Scenario 1 illustrates this risk: An ED capacity algorithm embedding neighborhood-based disparities.

Third, proprietary AI systems may shield decision-making from institutional oversight through intellectual property protections (Scenario 3). When vendors invoke "algorithmic trade secrets," hospitals cannot audit decision processes, creating governance blind spots incompatible with public accountability expectations.<sup>[30]</sup>



**Figure 1.** A governance framework for managerial artificial intelligence in hospitals.

### Authority Redistribution and Organizational Hierarchy

Managerial AI alters traditional authority patterns within hospital hierarchies. When algorithmic systems gain influence over strategic decisions, authority shifts from established organizational roles (executives, clinical leaders) toward technical experts who design and maintain AI systems.<sup>[31,32]</sup> This redistribution raises concerns about governance.

Data scientists and IT specialists – traditionally supporting rather than leading organizational strategy – acquire informal power through their unique ability to interpret algorithmic outputs. Yet these individuals may lack training in healthcare ethics, organizational strategy, or stakeholder engagement and are not elected or appointed to governance roles with public accountability.<sup>[33,34]</sup>

Simultaneously, executives and board members may experience deskilling – diminished capacity for independent judgment as they increasingly defer to algorithmic recommendations. Over-reliance on AI systems can erode institutional expertise necessary for effective governance, particularly in uncertain or unprecedented situations where algorithmic predictions may be unreliable.<sup>[35,36]</sup>

### A Governance Framework for Managerial AI in Hospitals

Addressing the governance challenges identified above requires frameworks explicitly designed for managerial AI contexts. We propose a three-dimensional framework linking managerial AI to decision authority, accountability mechanisms, and institutional legitimacy (Fig. 1).

The three-dimensional structure reflects distinct governance functions in accountability literature.<sup>[20,37,38]</sup> Dimension 1 (decision authority) addresses the “who decides” question central to organizational hierarchy and responsibility assignment. Dimension 2 (transparency) addresses the “how justifiable” question, essential to oversight and accountability. Dimension 3 (legitimacy) addresses the “why acceptable” question linking technical systems to organizational mission and stakeholder trust. These dimensions are analytically distinct yet interdependent: authority without transparency enables unaccountable decisions; transparency without authority creates information overload without responsibility; legitimacy without both authority and transparency becomes performative rather than substantive. This tripartite structure aligns with Bovens’ accountability framework<sup>[10]</sup> while adapting it specifically for AI-enabled organizational contexts.

### **Dimension 1: Explicit Decision Authority Assignment**

Principle: Organizations must explicitly assign decision authority when AI systems inform strategic, operational, or resource allocation decisions. Authority assignment clarifies who makes final decisions, who provides input, and who bears responsibility for outcomes.

#### **Implementation Mechanisms**

- Decision authority matrix: Specify decision rights, designating decision-makers, advisors, technical experts, and oversight
- Board-level oversight: Establish a governance committee with diverse expertise that meets quarterly
- Human-in-command: Require active approval of AI recommendations for critical decisions.

### **Dimension 2: Algorithmic Transparency and Explainability**

Principle: Organizations must ensure managerial AI systems are sufficiently transparent to enable meaningful oversight, stakeholder understanding, and accountability assignment. Transparency requirements differ from clinical explainability – emphasizing decision processes, data provenance, and governance traceability rather than individual prediction justification.

#### **Implementation Mechanisms**

- Impact assessments: Document system purpose, methodology, assumptions, and equity implications
- Vendor transparency: Require documentation of models, performance metrics, and data sources.

Implementation faces a critical challenge: Power asymmetry between hospitals and technology corporations. Individual hospitals often lack the bargaining power to demand algorithmic disclosure from vendors that invoke trade secrets. This requires regulatory intervention through: (1) national procurement standards mandating baseline thresholds for transparency for public healthcare AI systems; (2) consortium-based procurement where hospital groups collectively negotiate transparency clauses; and (3) regulatory frameworks requiring algorithmic impact assessments as market authorization conditions. Without such structural interventions, vendor transparency remains aspirational.

- Performance monitoring: Monitor equity metrics, conduct annual audits for high-stakes applications.

### **Dimension 3: Institutional Legitimacy Protection**

Principle: Hospitals must protect institutional legitimacy by ensuring that AI-supported decisions align with the organizational mission, professional values, and public accountability expectations. Legitimacy requires demonstrating that algorithmic systems serve patient and community interests rather than merely optimizing operational metrics.<sup>[34,35]</sup>

#### **Implementation Mechanisms**

- Stakeholder engagement processes: Engage stakeholders (clinical staff, patient councils, community representatives) to assess acceptability and incorporate diverse perspectives.
- Mission alignment reviews: Require governance committees to periodically assess whether AI applications align with organizational mission and values. Systems that optimize financial performance at the expense of access, equity, or quality should be modified or discontinued, regardless of technical sophistication.
- Public accountability mechanisms: Publicly report AI governance annually, including systems in use, governance structures, equity monitoring, and problem responses.

### **Discussion**

The framework's relevance is heightened by recent regulatory and policy developments, including the European Union's AI Act and Organization for Economic Cooperation and Development analysis on AI in health.<sup>[39,40]</sup> The World Health Organization has also issued guidance on ethics and governance of AI for health, including large multi-modal models.<sup>[41]</sup> While our review focuses on predictive and analytical AI, GenAI systems introduce additional governance concerns (e.g., non-deterministic outputs and hallucinations) that can affect managerial decision quality and therefore warrant supplementary controls and assurance mechanisms.<sup>[42]</sup>

### **Theoretical Contributions**

This review advances health sciences scholarship by theorizing managerial AI as a governance challenge. While existing literature addresses clinical AI ethics,<sup>[1–4,23,24]</sup> organizational governance implications remain underexplored. We contribute by: (1) distinguishing managerial from clinical AI across governance dimensions; (2) identifying accountability gaps as central challenges; and (3) proposing an integrated framework.

Our framework synthesizes organizational theory,<sup>[26,37,38,43]</sup> healthcare management,<sup>[5,6,44]</sup> and AI governance literature,<sup>[32–35,37,38,45]</sup> to address hospital-specific challenges. In contrast to the extensive citation of cross-sector AI ethics frameworks, which predominantly emphasize high-level principles such as fairness, transparency, and accountability, our approach is distinctive in its explicit focus on institutional accountability mechanisms within hospital organizations that are bound by public-interest obligations.

### Implications for the Turkish Healthcare System

The proposed framework is of relevance to the healthcare system of Türkiye, where the combination of centralized management, performance-driven incentives, and rapid digitalization gives rise to both opportunities and governance risks regarding the adoption of managerial AI.

The framework's relevance extends to Türkiye's rapidly evolving healthcare system, where hospital governance faces distinct structural and institutional challenges. Türkiye's Health Transformation Program (2003–ongoing) has emphasized performance measurement, efficiency optimization, and data-driven management, thereby creating a fertile institutional environment for the adoption of managerial AI.<sup>[22,25]</sup> However, effective implementation of the framework in the Turkish context requires explicit attention to several context-specific governance considerations.

**Centralized governance architecture.** Public hospitals in Türkiye operate within highly centralized governance structures spanning the Ministry of Health, provincial health directorates, and individual hospital administrations. This multi-layered arrangement necessitates formal coordination and escalation mechanisms that are often less pronounced in decentralized health systems. Consequently, clear and explicit assignment of decision authority (Framework Dimension 1) becomes particularly critical when AI-supported decision rights and accountability span multiple institutional levels.

### Performance-Based Management Culture

Türkiye's performance-based payment system creates strong AI-optimization incentives but may overshadow equity considerations for geographically isolated or disadvantaged populations. Legitimacy protection mechanisms (Framework Dimension 3) must ensure algorithmic optimization serves social mission alongside efficiency goals.

### Academic Medicine Integration

University hospitals play dual roles as care delivery organizations and research institutions. Managerial AI governance in teaching hospitals must balance operational efficiency with the educational mission and research priorities – requiring governance structures that accommodate multiple stakeholder perspectives and accountability relationships not fully captured by corporate governance models.<sup>[36,44]</sup>

### Policy Recommendations for Turkish Healthcare

We propose policy actions for Türkiye: (1) national AI governance standards with equity audits; (2) algorithmic fairness in performance-based payment; (3) mandatory impact assessments; (4) AI governance certification; and (5) consortium-based procurement.

**Data governance infrastructure:** Türkiye's National Health Information System provides a centralized health data infrastructure, enabling large-scale AI applications but raising data governance questions. Hospital-level managerial AI must navigate relationships with national data platforms, requiring transparency protocols (Framework Dimension 2) clarifying data provenance, access controls, and algorithmic auditability across organizational boundaries.

These considerations suggest that effective AI governance in Türkiye's hospitals requires context-sensitive adaptation of the framework to account for the centralized health system architecture, the longstanding emphasis on performance management, and the practical realities of data availability, interoperability, and infrastructure capacity. The proposed framework provides a conceptual foundation; implementation requires context-specific operationalization tailored to Turkish governance ecosystems.

### Practical Implications for Hospital Leaders

While managerial AI deployment remains limited in Türkiye, the framework's anticipatory value lies in establishing governance before widespread implementation, when practices are forming, and correction is less costly.

For hospital executives and governing boards, this framework suggests actionable strategies:

- Before AI adoption: Conduct algorithmic impact assessments, establish governance committees, and develop procurement criteria that emphasize transparency.

- During implementation: Create decision authority matrices, implement human-in-command protocols, and engage affected stakeholders.
- Ongoing governance: Monitor performance across equity metrics, conduct annual audits, and publicly report governance activities.
- Implementation requires governance coordination (0.2–0.5 FTE), technical expertise, and quarterly board reviews. Smaller hospitals may use shared committees. Governance investment prevents ungoverned AI risks.

#### Resource Considerations

- Governance effectiveness metrics: Decision authority clarity, transparency, compliance, stakeholder trust, equity monitoring, and accountability responsiveness.
- Phased implementation: Months 1–3: Governance committee, systems inventory, authority matrix; 4–6: Impact assessments, vendor transparency; 7–9: Monitoring, stakeholder engagement; 10–12: Annual audit, process refinement.

### Phased Implementation Roadmap

To support practical adoption, we provide operational guidance for implementing the governance framework. We align the recommended governance functions with the NIST AI Risk Management Framework (AI RMF 1.0) and the NIST Generative AI Profile to support risk-based implementation and assurance.<sup>[42,46]</sup> For organizations seeking a certifiable management-system approach, ISO/IEC 42001 can be used to formalize policies, roles, documentation, and continual improvement cycles around AI governance.<sup>[47]</sup> While organizations must adapt these mechanisms to their specific contexts, the templates and timelines offer actionable starting points.

### Implementation Guidance and Operationalization

Translating the proposed governance framework into practice requires systematic and staged operationalization aligned with organizational readiness and AI maturity. Rather than 1-time compliance activities, effective implementation should be understood as a progressive institutionalization process, in which governance capabilities are incrementally developed, reinforced, and embedded within routine managerial practice.

For Dimension 1 (Decision Authority), hospitals should establish explicit decision-authority matrices that clarify which AI-supported decisions require executive approval, board oversight, or escalation. Dedicated board-level or executive AI governance committees play a critical role in

maintaining accountability and ensuring that algorithmic recommendations do not displace human judgment.

For Dimension 2 (Transparency), organizations should conduct algorithmic impact assessments, define baseline thresholds for explainability, and require comprehensive documentation from internal teams and external vendors. These measures are essential to enable meaningful oversight and retrospective review.

For Dimension 3 (Institutional Legitimacy), proactive stakeholder engagement – including clinicians and managers, and, where appropriate, the public – should be complemented by transparent reporting practices that communicate how AI systems influence organizational decisions.

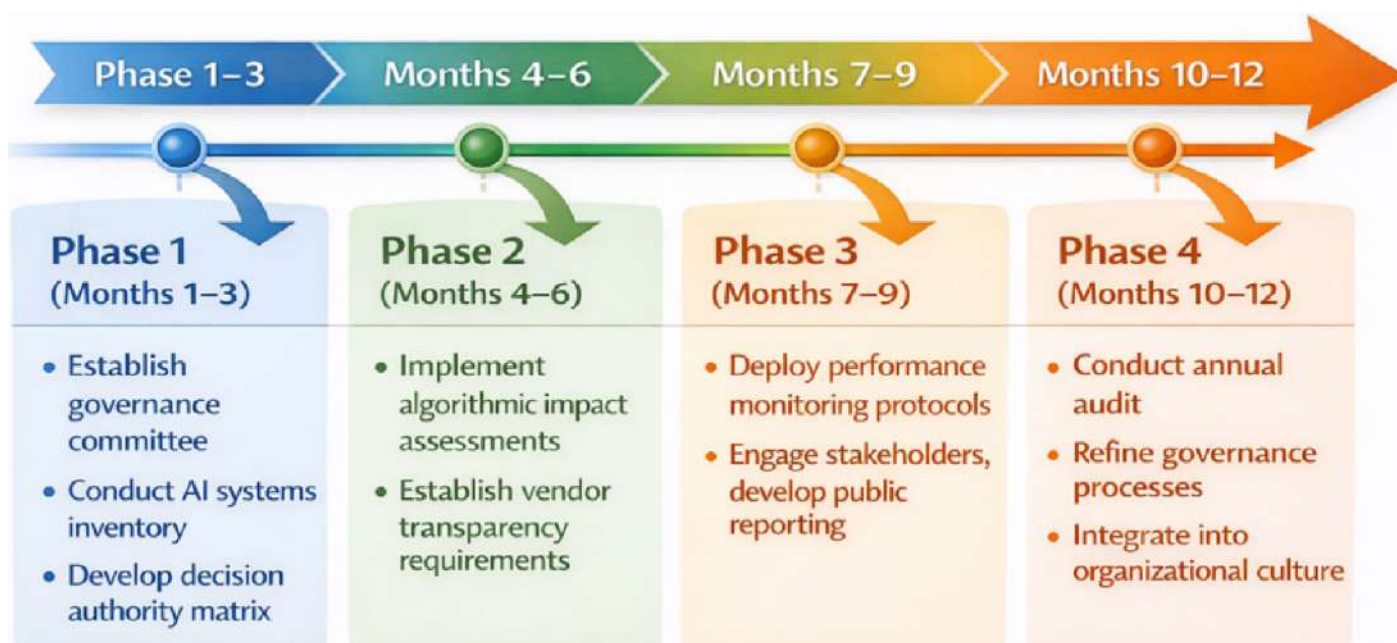
Figure 2 illustrates a 12-month, phased implementation roadmap for managerial AI governance in hospital organizations. The roadmap draws on operational lessons from an institutional responsible-AI guideline case study,<sup>[48]</sup> enterprise governance considerations for scaling AI in healthcare organizations,<sup>[49]</sup> and a prescriptive health-system evaluation and monitoring framework for the appropriate implementation and review-AI.<sup>[50]</sup> The phases are cumulative, such that activities initiated in earlier stages continue and mature throughout subsequent phases. This approach emphasizes sustained capability building – from foundational governance structures to performance monitoring and cultural integration – rather than discrete or isolated interventions. Hospitals should tailor the pace and sequencing based on their governance capacity, organizational readiness, and AI maturity levels.

Governance effectiveness across the implementation phases should be monitored using complementary indicators, including:

- Decision-authority clarity (proportion of AI-supported decisions with documented authority and escalation pathways);
- Transparency compliance (proportion of systems with completed impact and explainability assessments);
- Stakeholder trust (survey-based measures among clinicians, managers, and staff);
- Equity monitoring (regularly documented demographic impact reviews); and
- Accountability responsiveness (time from issue identification to corrective action).

### Limitations

Several limitations of the present study should be acknowledged when interpreting the findings and considering directions for future research.



**Figure 2.** Phased implementation roadmap for managerial artificial intelligence governance in hospital organizations.

First, the proposed governance framework has not yet been empirically validated. Although its theoretical foundations are well grounded in organizational theory, healthcare management, and AI governance literature, empirical testing remains necessary. Future research should validate the framework through rigorous methods, including expert consensus processes (e.g., Delphi studies), pilot implementations in hospital settings, and the development and testing of validated governance instruments. The guidance offered here should therefore be regarded as preliminary and subject to contextual adaptation.

Second, the narrative review methodology does not provide systematic evidence aggregation characteristic of systematic reviews. While narrative synthesis is appropriate for emerging and interdisciplinary domains, the interpretive nature of this approach may reflect author perspectives despite efforts to minimize bias. Future systematic reviews and meta-analyses could quantify the prevalence of specific governance challenges and assess the effectiveness of different managerial AI governance interventions.

Third, the framework is primarily oriented toward medium-to-large hospitals with established governance infrastructure. Smaller or resource-constrained organizations may lack the institutional capacity required for full implementation. In such contexts, governance mechanisms should be proportionate to organizational size, complexity, and AI maturity rather than uniformly applied.

Fourth, the rapid pace of AI development may outstrip existing governance frameworks. Emerging architectures, deployment models, and data practices may introduce governance challenges not fully anticipated in the present analysis. Effective AI governance should therefore be understood as a dynamic and adaptive process, requiring periodic revision as technologies and organizational uses evolve.

Fifth, although the framework is contextualized with reference to the Turkish healthcare system, its applicability across diverse governance models, regulatory environments, and cultural contexts remains to be established. Comparative and cross-national research would strengthen understanding of how managerial AI governance frameworks should be adapted to different institutional settings.

Despite these limitations, the framework offers actionable governance guidance by integrating insights from organizational theory, healthcare management, and AI governance literature. It provides a conceptual foundation for future empirical research and a practical starting point for institutional capacity-building in managerial AI governance.

## Conclusion and Future Research Directions

AI is transforming hospital management, yet governance frameworks lag adoption. This review repositions managerial AI as an institutional governance challenge that requires explicit attention to decision authority, accountability, and the protection of legitimacy.

The proposed framework addresses accountability gaps through three governance dimensions: Explicit assignment of decision authority, algorithmic transparency requirements, and protection of institutional legitimacy. These dimensions provide a conceptual foundation for hospitals navigating AI-enabled organizational transformation while maintaining public trust and accountability.

Future research should test the framework's implementation, develop validated instruments, and conduct comparative studies that engage diverse stakeholders.

As managerial AI becomes increasingly central to hospital operations, governance frameworks that ensure accountability, transparency, and legitimacy will determine whether algorithmic decision support serves the organizational mission and societal interests. This review provides theoretical foundation and practical guidance for that essential governance work.

**Ethics Committee Approval:** Ethical approval was not required for this study since this is a review article.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** The authors acknowledge the use of AI-assisted writing tools for language refinement, structural organization, and reference formatting during manuscript preparation. All conceptual framing, analytical arguments, and substantive content remain the original intellectual contribution of the authors.

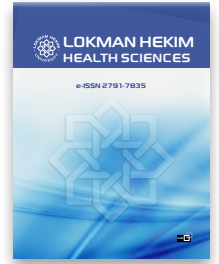
**Authorship Contributions:** Concept: İŞ, FDZ; Design: İŞ, FDZ; Supervision: İŞ, FDZ; Analysis and/or interpretation: İŞ, FDZ; Literature review: İŞ, FDZ; Writing: İŞ; Critical review: İŞ, FDZ.

**Peer-review:** Double blind peer-reviewed.

## References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44-56. [\[CrossRef\]](#)
2. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med* 2019;25(1):24-9. [\[CrossRef\]](#)
3. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med* 2018;378(11):981-3. [\[CrossRef\]](#)
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380(14):1347-58. [\[CrossRef\]](#)
5. Shortell SM, Kaluzny AD. Health care management: Organization Design and Behavior. 7<sup>th</sup> ed. Boston: Cengage Learning; 2020.
6. Hulshof PJH, Kortbeek N, Boucherie RJ, Hans EW, Bakker PJM. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Syst* 2012;1:129-75. [\[CrossRef\]](#)
7. Eeckloo K, Van Herck G, Van Hulle C, Vleugels A. From corporate governance to hospital governance. Authority, transparency and accountability of Belgian non-profit hospitals' board and management. *Health Policy* 2004;68(1):1-15. [\[CrossRef\]](#)
8. Alexander JA, Lee SY, Bazzoli GJ. Governance forms in health systems and health networks. *Health Care Manage Rev* 2003;28(3):228-42. [\[CrossRef\]](#)
9. Saltman RB, Ferroussier-Davis O. The concept of stewardship in health policy. *Bull World Health Organ* 2000;78(6):732-9.
10. Bovens M. Analysing and assessing accountability: a conceptual framework. *Eur Law J* 2007;13(4):447-68. [\[CrossRef\]](#)
11. Diakopoulos N. Accountability in algorithmic decision making. *Commun ACM* 2016;59(2):56-62. [\[CrossRef\]](#)
12. Green BN, Johnson CD, Adams A. Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. *J Chiropr Med* 2006;5(3):101-17. [\[CrossRef\]](#)
13. Cronin P, Ryan F, Coughlan M. Undertaking a literature review: a step-by-step approach. *Br J Nurs* 2008;17(1):38-43. [\[CrossRef\]](#)
14. Coombs C, Hislop D, Taneva SK, Barnard S. The strategic impacts of intelligent automation for knowledge and service work: an interdisciplinary review. *J Strateg Inf Syst* 2020;29:101600. [\[CrossRef\]](#)
15. Hassan M, Borycki EM, Kushniruk AW. Artificial intelligence governance framework for healthcare. *Healthc Manage Forum* 2025;38(2):125-30. [\[CrossRef\]](#)
16. Donia J. Algorithmic accountabilities and health systems: a review and sociomaterial approach. *Big Data Soc* 2025;12(2):20539517251334099. [\[CrossRef\]](#)
17. Bartsch S, Behn O, Benlian A, Brownsword R, Bücken S, Düwell M, et al. Governance of high-risk AI systems in healthcare and credit scoring. *Bus Inf Syst Eng* 2025;67:563-81. [\[CrossRef\]](#)
18. Raisch S, Krakowski S. Artificial intelligence and management: the automation-augmentation paradox. *Acad Manage Rev* 2021;46(1):192-210. [\[CrossRef\]](#)
19. Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. *Technol Forecast Soc Change* 2018;126:3-13. [\[CrossRef\]](#)
20. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. *Minds Mach (Dordr)* 2018;28(4):689-707. [\[CrossRef\]](#)
21. Matthias A. The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf Technol* 2004;6(3):175-83. [\[CrossRef\]](#)
22. Pasquale F. *The Black Box Society: The secret algorithms that control money and information*. Cambridge: Harvard University Press; 2015. [\[CrossRef\]](#)
23. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;1(5):206-15. [\[CrossRef\]](#)

24. Burrell J. How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc* 2016;3(1):2053951715622512. [CrossRef]
25. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366(6464):447-53. [CrossRef]
26. Mintzberg H. The structure of organizations. In: Asch D, Bowman C, editors. *Readings in strategic management*. London: Palgrave; 1989. p. 322-52. [CrossRef]
27. Barocas S, Selbst AD. Big data's disparate impact. *Calif Law Rev* 2016;104:671-732. [CrossRef]
28. Lebovitz S, Levina N, Lifshitz-Assaf H. Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Q* 2021;45(3):1501-26. [CrossRef]
29. Eubanks V. *Automating inequality: How high-tech tools profile, police and punish the poor*. New York: St. Martin's Press; 2018.
30. Brynjolfsson E, Mitchell T. What can machine learning do? Workforce implications. *Science* 2017;358(6370):1530-4. [CrossRef]
31. European Commission High-Level Expert Group on AI. *Ethics guidelines for trustworthy AI*. Brussels: European Commission; 2019.
32. Reisman D, Schultz J, Crawford K, Whittaker M. *Algorithmic impact assessments: a practical framework for public agency accountability*. New York: AI Now Institute; 2018.
33. Morley J, Machado CCV, Burr C, Cows J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. *Soc Sci Med* 2020;260:113172. [CrossRef]
34. Suchman MC. Managing legitimacy: strategic and institutional approaches. *Acad Manage Rev* 1995;20(3):571-610. [CrossRef]
35. Deephouse DL, Suchman M. Legitimacy in organizational institutionalism. In: Greenwood R, Oliver C, Lawrence TB, Meyer RE, eds. *The SAGE Handbook of Organizational Institutionalism*. 2<sup>nd</sup> ed. London: SAGE; 2017:27-54. [CrossRef]
36. National Academy of Medicine. *Vital Directions for Health & Health Care: An Initiative of the National Academy of Medicine*. Washington, DC: The National Academies Press; 2017.
37. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc* 2020;27(3):491-7. [CrossRef]
38. Emanuel EJ, Emanuel LL. What is accountability in health care? *Ann Intern Med* 1996;124(2):229-39. [CrossRef]
39. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union* 2024;L(2024/1689). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
40. Organization for Economic Co-operation and Development (OECD). *AI in Health: Huge Potential, Huge Risks*. OECD Artificial Intelligence Papers, No. 9. Paris: OECD Publishing; 2024. [https://www.oecd.org/en/publications/2024/01/ai-in-health-huge-potential-huge-risks\\_ff823a24.html](https://www.oecd.org/en/publications/2024/01/ai-in-health-huge-potential-huge-risks_ff823a24.html)
41. World Health Organization (WHO). *Ethics and governance of artificial intelligence for health: guidance on large multi-modal models (LMMs)*. Geneva: WHO; 2025. <https://www.who.int/publications/i/item/9789240084759>
42. National Institute of Standards and Technology (NIST). *Artificial intelligence risk management framework: generative artificial intelligence profile*. NIST AI 600-1. Gaithersburg (MD): NIST; 2024 Jul. [CrossRef]
43. Simon HA. *Administrative Behavior*. 4<sup>th</sup> ed. New York: Free Press; 1997.
44. Glouberman S, Mintzberg H. Managing the care of health and the cure of disease--Part I: Differentiation. *Health Care Manage Rev* 2001;26(1):56-69; discussion 87-9. [CrossRef]
45. Brinkerhoff DW. Accountability and health systems: toward conceptual clarity and policy relevance. *Health Policy Plan* 2004;19(6):371-9. [CrossRef]
46. National Institute of Standards and Technology (NIST). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg (MD): NIST; 2023. <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf>
47. International Organization for Standardization / International Electrotechnical Commission. *ISO/IEC 42001:2023 Artificial intelligence - Management system*. Geneva: ISO; 2023. <https://www.iso.org/standard/42001.html>
48. Saenz AD, Centi A, Ting D, You JG, Landman A, Mishuris RG. Establishing responsible use of AI guidelines: a comprehensive case study for healthcare institutions. *NPJ Digit Med* 2024;7(1):348. [CrossRef]
49. Bodnari A, Travis J. Scaling enterprise AI in healthcare: the role of governance in risk mitigation frameworks. *NPJ Digit Med* 2025;8(1):272. [CrossRef]
50. Wells BJ, Nguyen HM, McWilliams A, Pallini M, Bovi A, Kuzma A, et al. A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare. *NPJ Digit Med* 2025;8(1):514. [CrossRef]



## REVIEW

# Integrating Artificial Intelligence into Assisted Reproduction Technologies: Current Applications and Future Directions – A Narrative Review

Rüyam Ercenk, Suna Yıldırım Karaca

Department of Obstetrics and Gynecology, İzmir Bakırçay University Faculty of Medicine, İzmir, Türkiye

## Abstract

In recent years, the integration of artificial intelligence (AI) into medicine has expanded rapidly, particularly within assisted reproductive technologies (ART) and *in vitro* fertilization (IVF). Traditional assessments in IVF—especially embryo morphology—are prone to subjectivity and may vary according to embryologist experience. AI-supported systems help overcome these limitations by enabling faster, more objective, and more consistent evaluation of clinical data and microscopic images. AI applications have been incorporated into multiple steps of the ART process, including oocyte and sperm assessment, fertilization analysis, embryo evaluation, ploidy prediction, and embryo selection for transfer. Beyond laboratory assessment, AI also contributes to micromanipulation, quality management, the processing of large datasets to support personalized treatment protocols, and improved genetic testing approaches. Collectively, these innovations enhance diagnostic accuracy, promote standardization, and increase treatment success rates in ART. This narrative review provides a comprehensive and up-to-date overview of AI applications within ART, with a particular focus on IVF laboratory processes, clinical decision-support tools, and related ethical considerations. A focused literature search was conducted in PubMed using the keywords “artificial intelligence” and “assisted reproduction.” The search covered the period from January 1, 2020, to May 31, 2025, and included only English- and Turkish-language publications. Eligible studies consisted of meta-analyses, systematic reviews, narrative reviews, and original research evaluating the use of AI in human ART or IVF. Conference abstracts, editorials, expert opinions, letters to the editor, case reports lacking methodological clarity, non-human studies, and purely technical computer science papers without clinical relevance were excluded. Reference lists of the included articles were also examined to identify additional sources.

**Keywords:** Artificial intelligence; Assisted reproductive technologies; Individualized medicine; *in vitro* fertilization

**I**nfertility is a disease of the male or female reproductive system and is defined as the failure to achieve pregnancy after 12 months or more of regular, unprotected sexual intercourse. Infertility may result from male, female, or

unexplained factors.<sup>[1]</sup> Approximately 17% of people experience infertility at some point in their lives. Prevalence estimates are consistent across countries regardless of income level; 17.8% of individuals in high-income countries

**Cite this article as:** Ercenk R, Yıldırım Karaca S. Integrating Artificial Intelligence Into Assisted Reproduction Technologies: Current Applications and Future Directions – A Narrative Review. Lokman Hekim Health Sci 2026;6(2):352–359.

**Correspondence:** Rüyam Ercenk, M.D. İzmir Bakırçay Üniversitesi Tıp Fakültesi, Kadın Hastalıkları ve Doğum Anabilim Dalı, İzmir, Türkiye  
**E-mail:** ruyamercken@gmail.com **Submitted:** 08.08.2025 **Revised:** 30.11.2025 **Accepted:** 17.02.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



and 16.5% in low- and middle-income countries experience infertility. The prevention, diagnosis, and treatment of infertility remain underfunded and inaccessible for many individuals.<sup>[2]</sup> In Türkiye, the prevalence of infertility ranges between 10% and 20%, with 55–75% of these couples diagnosed with primary infertility and 25–40% with secondary infertility.<sup>[3]</sup>

The development of assisted reproductive technologies (ART) began in 1978 with the first successful *in vitro* fertilization (IVF), marking a groundbreaking milestone in the field. In the following years, alternative methods such as GIFT (Gamete Intrafallopian Transfer) and ZIFT (Zygote Intrafallopian Transfer) were introduced, expanding treatment options for infertility. In the 1990s, innovative techniques such as intracytoplasmic sperm injection (ICSI)—a revolutionary advancement particularly in the treatment of male infertility—and preimplantation genetic diagnosis (PGD), which enables screening of embryos for genetic disorders, were incorporated into clinical practice. During the 2000s, significant improvements in embryology laboratory techniques and cryopreservation methods—especially the widespread adoption of vitrification—led to substantial increases in success rates. In the following years, the effectiveness and scope of ART were further enhanced by the introduction of controlled ovarian hyperstimulation, luteal phase support, embryonic genetic testing, and oocyte cryopreservation. Today, novel approaches such as minimal stimulation protocols, GnRH agonist cycle triggers, and metabolomic/proteomic analyses continue to shape ART as an increasingly evolving and personalized field.

Conventional ART methods, such as *in vitro* fertilization and intracytoplasmic sperm injection, face several challenges that limit their effectiveness. These challenges include low success rates, high costs, invasive procedures, ethical concerns, a lack of personalized approaches, and the inherently subjective nature of both embryology and clinical practice. All these factors highlight the need for more standardized and technology-driven approaches to enhance the reliability and consistency of ART procedures.<sup>[4]</sup> Therefore, the aim of this narrative review is to provide a comprehensive and up-to-date synthesis of current artificial intelligence applications within assisted reproductive technologies, with a particular focus on IVF laboratory processes, clinical decision-support tools, and the associated ethical and regulatory considerations.

This narrative review was conducted using a structured literature search in the PubMed database. The search strategy combined the keywords “artificial intelligence”

AND “assisted reproduction.” The primary search covered publications from January 1, 2020, to May 31, 2025. Articles written in English or Turkish were included. Eligible publications comprised meta-analyses, systematic reviews, narrative reviews, and original research articles evaluating the use of artificial intelligence in human assisted reproductive technology or *in vitro* fertilization settings. Non-human studies, case reports lacking methodological clarity, editorials, expert opinions, letters to the editor, and purely technical computer science reports without clinical or laboratory relevance to assisted reproduction were excluded. Reference lists of the included publications were also screened to identify earlier seminal studies, relevant methodological studies, and selected preliminary reports that directly supported specific technical or laboratory statements. Because this was designed as a narrative review, no formal risk-of-bias assessment or meta-analysis was performed.

## Clinical and Research Consequences

### The Use of Artificial Intelligence in Assisted Reproductive Technologies

Today, artificial intelligence (AI) applications are widely used across various fields of medicine, offering significant advantages, particularly in areas such as medical imaging, personalized treatment planning, and robotic surgery. As a reflection of these advancements, the integration of AI-based approaches into the field of ART is bringing important innovations to reproductive medicine and gradually reshaping clinical and laboratory practices.<sup>[5]</sup>

This review provides a comprehensive overview of the integration of AI into ART, focusing on oocyte assessment and selection, sperm evaluation and selection, and embryo assessment and selection. It examines in detail the impact of AI on personalized treatment planning, diagnostic accuracy, ethical considerations, and the future of this rapidly evolving field.

### Artificial Intelligence in the *in Vitro* Fertilization Laboratory

Manual procedures still dominate many steps of the *in vitro* fertilization laboratory. However, the integration of automation and AI has significant potential to reduce subjectivity in embryology, thereby increasing consistency and minimizing variability between embryologists.<sup>[6]</sup> In addition, AI-based sensors, processors, and integrated software systems enable real-time monitoring of key environmental parameters such as room temperature,

humidity, volatile organic compounds, and door-opening frequency, and can initiate automated responses when deviations occur.<sup>[7]</sup> In IVF laboratories, AI-driven automation enhances workflow efficiency by performing continuous and objective assessments of gametes and embryos through time-lapse imaging platforms and deep-learning-based image analysis.<sup>[8]</sup> These algorithms automatically annotate morphokinetic events, detect fertilization patterns, track cell divisions, and flag potential errors in embryo identification, thereby reducing inter-observer variability and improving standardization.<sup>[8,9]</sup> Furthermore, AI-supported systems contribute to quality management by monitoring incubator conditions, equipment performance, and overall laboratory workflow, allowing early detection of deviations and reducing the risk of technical errors.<sup>[7,9]</sup>

### Personalization of Ovarian Stimulation Protocols

The management of *in vitro* fertilization cycles depends on the ovarian response to treatment, which is evaluated during regular follow-ups to enable clinicians to make informed decisions and plan subsequent steps of the therapy. Selecting a personalized gonadotropin dose for oocyte stimulation is a complex process. Currently, the approach to determining the appropriate gonadotropin dose relies largely on the clinician's experience and the patient's response to previous treatment attempts.<sup>[10]</sup> During the ovarian stimulation phase, AI algorithms collect and process demographic characteristics, medical history, laboratory test results, medication dosages during treatment, and follicular ultrasound measurements to predict critical clinical decisions made by physicians, such as continuation of stimulation, timing of monitoring, dose adjustment, or cycle cancellation.

AI research on the selection of ovarian stimulation protocols remains limited, with only a small number of models developed to date.<sup>[6]</sup> Most published studies rely on retrospective, single-center datasets with heterogeneous samples, which restricts both reproducibility and generalizability.<sup>[11,12]</sup> In contrast to embryo assessment—where AI models have been trained on tens of thousands of images—AI-driven stimulation models are based on relatively small patient cohorts and lack external validation.<sup>[6,9]</sup> Moreover, no completed randomized controlled trials have yet demonstrated the clinical superiority of AI-guided protocols, underscoring the need for larger, prospectively designed studies before these tools can be routinely implemented in clinical decision-making.<sup>[6]</sup>

In the study by Correa et al.,<sup>[11]</sup> an AI model was developed to predict individualized starting doses of FSH for ovarian

stimulation, using variables such as age, body mass index, anti-Müllerian hormone, antral follicle count, and previous live birth history. The performance scores obtained during the development and validation phases were found to be statistically significantly higher compared with the doses prescribed by clinicians. This AI-based approach not only supports the decision-making process and quality control monitoring but also has the potential to improve treatment efficiency by reducing the risk of cycle cancellation through personalized dose recommendations.

Fanton et al.<sup>[12]</sup> developed an AI algorithm based on the principle of patient similarity to optimize the selection of the initial FSH dose. The model generated personalized dose–response curves based on baseline parameters such as age, body mass index, basal AMH, and antral follicle count, thereby identifying the optimal dose range. Using this model, a higher number of fertilized embryos and usable blastocysts was obtained, while total FSH consumption was significantly reduced. These findings demonstrate the potential of AI-based dose recommendation systems to improve clinical outcomes and reduce costs.

### Oocyte Assessment and Selection

The assessment of oocyte quality is a critical step in ART, as oocyte competence strongly influences subsequent embryo development and overall IVF success. Follicular volume correlates with the degree of ovarian response to stimulation and is widely recognized as an indirect marker of oocyte maturation.<sup>[13,14]</sup>

Because follicular size and volume serve as key maturation indicators, transvaginal ultrasound naturally becomes the primary method for monitoring follicular development during stimulation. Although two-dimensional ultrasound provides baseline measurements, three-dimensional ultrasound offers superior volumetric assessment and has been associated with improved planning for oocyte retrieval.<sup>[13]</sup> However, both 2D and 3D measurements remain highly operator-dependent, and variability can occur both between different clinicians and within repeated assessments by the same individual.<sup>[14]</sup> These limitations underscore the need for more objective and automated assessment tools in oocyte evaluation.

Recent advances in AI-based image analysis have made significant contributions in this area. Deep-learning models using transvaginal ultrasound datasets have enabled automated segmentation and quantification of ovaries and follicles, potentially reducing operator dependence in follicular monitoring.<sup>[6]</sup>

Targosz et al.<sup>[15]</sup> showed that deep neural networks can be used for semantic segmentation of human oocyte images, enabling automated identification of relevant morphological structures and supporting more standardized oocyte image analysis. Similarly, Firuzinia et al.<sup>[16]</sup> developed a robust deep-learning–based multiclass segmentation method for analyzing human metaphase II oocyte images, enabling automated evaluation of key morphological structures. These advances collectively demonstrate how AI-driven systems can strengthen objectivity and standardization in oocyte assessment.<sup>[6]</sup>

### Timing of Trigger Injection

In ART, determining the optimal timing of the trigger injection is a complex decision-making process influenced by factors such as follicle size and hormonal indicators. However, AI, with its capacity to integrate multidimensional datasets, holds the potential to enhance the accuracy and precision of this decision-making step.<sup>[8]</sup>

In the study by Hariton et al.,<sup>[17]</sup> it was demonstrated that the use of machine learning algorithms to optimize the timing of the trigger injection in an IVF cycle could lead to a significant increase in both the number of fertilized oocytes and the total number of usable blastocysts compared with physician-determined decisions.

Letterie et al.<sup>[18]</sup> developed an AI-based algorithm aimed at predicting the optimal trigger day for oocyte retrieval during ovarian stimulation in the IVF process, using data from only a single monitoring day. The model was built upon pre-IVF clinical profiles, such as age, AMH level, and BMI, as well as estradiol levels, follicle count, and follicle diameters obtained from one day of the stimulation cycle. It was able to identify not only the ideal trigger day but also a three-day window—within a  $\pm 1$  day tolerance—for oocyte retrieval. The model had a mean prediction error of 1.355 days and tended to estimate the trigger day earlier than the actual clinical decision, providing flexibility for additional clinical evaluation if needed. Furthermore, it was reported that shifting the trigger day forward or backward by one day resulted in an average variation of 0 to 3 in the total number of oocytes retrieved. This approach has the potential to enhance clinical efficiency.

### Sperm Assessment and Selection

Male infertility is the primary cause in approximately half of all infertile couples.<sup>[19]</sup> In the context of IVF, sperm selection based on embryologist evaluation remains subjective, thereby contributing to interobserver variability.<sup>[20]</sup> Consequently, this underscores the need for objective

and standardized methods in sperm assessment. To address this, Ottl et al.<sup>[21]</sup> developed an AI algorithm that facilitates the selection of the fastest sperm for fertilization by calculating head movement velocity. Similarly, Riordon et al.<sup>[22]</sup> demonstrated that an AI algorithm analyzing sperm head morphology surpassed traditional assessment methods, achieving an accuracy of 94.1%. Furthermore, Sato et al.<sup>[23]</sup> introduced an algorithm capable of both tracking sperm and evaluating them morphologically. In addition, Mendizabal-Ruiz et al.<sup>[24]</sup> introduced SiD software, which enables real-time single-sperm selection by computing motility parameters, including straight-line velocity and the linearity of the curvilinear path. Moreover, increased DNA fragmentation in sperm cells is known to negatively affect fertilization potential.<sup>[25]</sup> In response to this challenge, McCallum et al.<sup>[26]</sup> introduced an AI-based algorithm capable of predicting DNA fragmentation in sperm cells using bright-field images, further expanding the scope of objective and non-invasive sperm assessment.

Despite these promising technical results, the clinical implementation of AI-based sperm assessment tools remains challenging. Most available algorithms are developed using heterogeneous, single-center datasets and require extensive validation before routine use. In addition, high costs, the need for standardized imaging conditions, and the absence of universally accepted laboratory protocols currently limit the widespread adoption of these systems in daily andrology practice.

### Fertilization Assessment

As the earliest measurable indicator of successful gamete interaction, fertilization assessment provides essential information for early embryo development and represents a key area in which AI may offer objective and clinically meaningful support.<sup>[6,8]</sup>

Dimitriadis et al.<sup>[9]</sup> developed an algorithm capable of distinguishing between normally and abnormally fertilized oocytes, achieving an accuracy of 93.1%.<sup>[27]</sup> These findings highlight the potential of AI in the assessment of fertilization. Although further research is needed, the use of AI in fertilization evaluation may contribute to higher pregnancy rates through more efficient embryo selection.

### Embryo Assessment

Manual evaluation of embryo morphology is inherently subjective and demonstrates substantial intra- and interobserver variability.<sup>[28]</sup> AI-assisted image analysis reduces this variability by more precisely identifying subtle morphological patterns, enabling more objective

and standardized embryo assessment. Consequently, AI-supported methods may improve the accuracy of identifying embryos with high implantation potential and may contribute to improved ART outcomes.<sup>[9]</sup>

AI algorithms trained on microscopic embryo images have been shown to accelerate and improve the accuracy of assessing developmental potential.<sup>[29,30]</sup> Khosravi et al.<sup>[31]</sup> developed a deep-learning platform trained with blastocyst images classified as good or poor quality by embryologists, achieving a fully automated workflow with 96% accuracy. These findings emphasize the potential of AI tools to reduce workload while enhancing consistency in embryo quality assessment.

Embryonic aneuploidy remains the leading cause of IVF failure, contributing to implantation failure and recurrent miscarriage.<sup>[32]</sup> Ploidy assessment via preimplantation genetic testing for aneuploidy (PGT-A) is widely used to identify euploid embryos for transfer. However, because PGT-A requires a trophectoderm biopsy, its invasive nature may introduce risks such as impaired embryo integrity and reduced implantation potential.<sup>[28]</sup>

Recent advances demonstrate that AI technologies can non-invasively predict embryo ploidy status using imaging and clinical data. Kato et al.<sup>[33]</sup> reported that existing AI-based, morphokinetic, and morphological embryo selection models showed associations with blastocyst euploidy rates. Likewise, Jiang et al.<sup>[34]</sup> integrated patient characteristics such as age, AMH levels, and sperm quality with blastocyst images, yielding significantly improved predictive accuracy.<sup>[9]</sup> These approaches offer the promise of selecting embryos with higher implantation potential without the risks associated with biopsy.<sup>[35]</sup>

However, the use of AI-based non-invasive ploidy prediction introduces important ethical considerations. False-positive predictions may result in the unjustified exclusion of potentially viable euploid embryos, whereas false-negative predictions could lead to the transfer of aneuploid embryos and associated adverse outcomes. These risks highlight the need for rigorous validation, transparent algorithmic reporting, and careful clinical oversight to ensure that AI-assisted embryo selection remains both safe and ethically responsible.

### **Prediction of Implantation Success**

As a critical determinant of treatment success in ART, predicting which embryo is most likely to implant enables clinicians to optimize embryo selection and improve overall pregnancy outcomes. In the study by Fitz et

al.,<sup>[36]</sup> which compared embryologists' performance in selecting day 5 euploid blastocysts with and without AI assistance, embryologists correctly selected the embryo that resulted in successful implantation 65.5% of the time without AI, whereas the accuracy increased to 73.6% with AI support. Across all participating embryologists, the average improvement in selecting the embryo with the highest implantation potential was 11.1% with AI assistance. Although no statistically significant differences were observed based on the level of experience, a more pronounced improvement trend was noted among less experienced embryologists.

### **Prediction of Clinical Outcomes and Designing Individualized Treatment Protocols**

AI is increasingly being utilized to predict clinical pregnancy outcomes in ART.<sup>[6]</sup> By integrating diverse clinical variables, including demographic factors, medical and reproductive history, cause of infertility, comorbidities, and previous ART attempts, AI models can estimate individualized success probabilities prior to treatment.<sup>[10]</sup> Goyal et al.<sup>[37]</sup> conducted a large-scale study evaluating AI models trained on 25 clinical parameters from 141,160 patients to predict the likelihood of live birth in IVF. The study demonstrated that excluding variables reduced model accuracy, emphasizing that the combined effect of multiple clinical features enhances predictive performance. These findings highlight the potential of AI as a clinical decision-support tool in estimating personalized treatment outcomes in IVF.

Building on these predictive capabilities, AI can also contribute to the design of individualized ovarian stimulation and treatment protocols. By analyzing patient-specific patterns within large datasets, AI systems can recommend tailored treatment strategies that may improve clinical efficiency and reduce the trial-and-error approach often observed in ART cycles. This personalized method aims to enhance treatment success while minimizing the burden associated with repeated unsuccessful attempts.

### **The Use of Micromanipulation and Robotic Applications in the ART Laboratory**

The use of robotic systems in embryo culture, monitoring, and cryopreservation contributes to automation and optimization in ART laboratories by reducing errors and improving record-keeping practices.<sup>[4]</sup> In a recently published groundbreaking case report, Mendizabal-Ruiz et al.<sup>[38]</sup> described a robot-assisted, digitally controlled, and remotely operated intracytoplasmic ICSI system and

reported the first successful live birth achieved using this platform. The robotic system autonomously performed approximately half of the 23 micromanipulation steps involved in ICSI, while the remaining procedures were executed remotely by an operator via digital commands. The platform integrates advanced technologies such as AI-assisted sperm selection, laser-based sperm immobilization, zona pellucida thinning using laser, and piezo-assisted injection. In clinical application, one of the blastocysts derived from oocytes fertilized via remote robotic ICSI resulted in a successful pregnancy and the birth of a healthy baby. This study represents the first clinical success demonstrating the feasibility of robotic automation in ART and marks a significant milestone in enhancing standardization and global accessibility of laboratory procedures.

### **Ethical Concerns Regarding the Use of Artificial Intelligence in Assisted Reproductive Technologies**

The integration of AI into ART necessitates not only clinical considerations but also a range of ethical evaluations. As these technologies operate directly on patient data, issues such as the protection of privacy, transparency in the informed consent process, and equitable access to services must be carefully addressed.<sup>[8]</sup>

In reproductive medicine, AI systems analyze individuals' genetic information, fertility-related medical history, and biological data. Therefore, ensuring the security of personal health information is not only a legal obligation but also an ethical imperative.<sup>[8]</sup> Data protection frameworks such as the European Union's General Data Protection Regulation (GDPR) and Türkiye's Personal Data Protection Law (KVKK) impose strict requirements for the processing of sensitive health data in clinical and laboratory workflows. Within this context, the anonymization or pseudonymization of datasets, restriction of third-party access, and the provision of transparent information regarding how AI systems use patient data during the informed consent process are of primary importance.

The inclusion of AI as a decision-support tool in clinical workflows does not absolve physicians of responsibility. Technological recommendations must always be interpreted alongside expert clinical judgment.<sup>[8]</sup> In cases in which AI-generated suggestions conflict with clinical assessment, final decisions must be made by healthcare professionals in accordance with medical ethical principles. Furthermore, in the event of adverse outcomes resulting

from AI-generated predictions, responsibility must be clearly defined—whether it lies with software developers, healthcare institutions, or physicians.

Laboratory-based risks also warrant attention, as insufficiently validated AI tools may increase the likelihood of misclassification, embryo misidentification, or inappropriate embryo selection, underscoring the need for rigorous quality control and oversight.<sup>[9]</sup> To ensure the ethical implementation of AI, a multidisciplinary approach is required. Effective collaboration among technology developers, healthcare professionals, legal experts, and ethics committees is necessary to guarantee that these systems are developed in compliance with both technical and ethical standards.<sup>[8,9]</sup>

To address these challenges, several regulatory and governance pathways can guide the safe integration of AI into ART. Agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have introduced frameworks emphasizing transparency, clinical validation, and post-market surveillance for AI-based medical technologies. Similarly, the Turkish Ministry of Health provides regulatory guidance through national medical device and digital health regulations. Implementing standardized reporting systems, external validation mechanisms, algorithmic audit procedures, and structured informed-consent models can help mitigate data-security risks, reduce algorithmic bias, and support the ethical and responsible adoption of AI-driven tools in reproductive medicine.

### **The Future of Artificial Intelligence in Assisted Reproductive Technologies and Potential Innovations**

In the future, AI is expected to become a fundamental tool for clinical decision support throughout the entire ART process—from ovarian stimulation to embryo transfer. This transformation will not only accelerate procedures but also promote the widespread adoption of personalized treatment plans that take into account individuals' genetic profiles, hormonal status, and environmental factors.<sup>[8]</sup> In procedures requiring micromanipulation, the reduction of human error is likely to positively influence success rates; moreover, the concept of AI-driven, fully automated embryology laboratories is becoming increasingly tangible.<sup>[4,8]</sup>

However, these technological advancements also raise important concerns. In particular, data security, algorithmic bias, ethical practice, and regulatory gaps stand out as

critical issues that shape the integration of AI into clinical practice.<sup>[5]</sup> If these challenges can be successfully addressed, AI-supported ART applications will not only enhance success rates but also transform reproductive medicine into a more personalized, predictable, and safer domain—ultimately paving the way for healthier pregnancies.<sup>[4,8]</sup>

## Conclusion

The use of AI in ART is rapidly expanding across both laboratory and clinical domains. Numerous AI-based algorithms have been developed not only for quality control in ART laboratories but also for enabling standardized and objective evaluations in steps traditionally dependent on embryologist expertise, such as oocyte quality assessment, sperm selection, ploidy prediction, embryo selection, and micromanipulation. By reducing subjectivity and improving reproducibility, these technologies contribute to more consistent decision-making and enhanced treatment efficiency. Moreover, AI-driven models show promise in predicting implantation success and generating personalized treatment protocols, offering opportunities to further refine clinical outcomes.

In conclusion, the integration of AI into infertility treatment through ART holds substantial potential to transform reproductive medicine. However, to translate this potential into routine clinical practice, more multifunctional and comprehensive models, expanded research efforts, large-scale randomized clinical trials, and interdisciplinary collaboration are needed.<sup>[4,6,9,29]</sup> It is equally important to recognize that current evidence is predominantly based on retrospective, single-center studies with limited sample sizes and a lack of randomized controlled trials, which restricts the generalizability of existing findings. Strengthening the evidence base through robust, well-designed studies will therefore be essential for the safe, reliable, and ethically responsible implementation of AI-assisted approaches in ART.

**Ethics Committee Approval:** Ethical approval was not required for this study since this is a review article.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Use of AI for Writing Assistance:** Not declared.

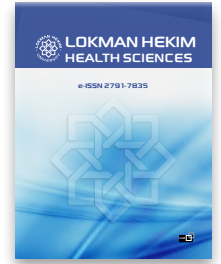
**Authorship Contributions:** Concept: RE; Design: RE; Supervision: SYK; Data collection and/or processing: RE; Analysis and/or interpretation: RE; Literature review: RE; Writing: RE; Critical review: SYK.

**Peer-review:** Double blind peer-reviewed.

## References

1. World Health Organization. Infertility [Internet]. Geneva: World Health Organization; 2024 May 22 [cited 2026 Apr 24]. Available from: <https://www.who.int/news-room/fact-sheets/detail/infertility>
2. Harris E. Infertility affects 1 in 6 people globally. *JAMA* 2023;329(17):1443. [\[CrossRef\]](#)
3. Aksu S, Sayan O. Determination of stigma, stress and depression levels in women receiving infertility treatment at a university hospital in the West Marmara region. *Istanbul Gelisim University Journal of Health Sciences* 2023;19:178-93. [\[CrossRef\]](#)
4. Mapari SA, Shrivastava D, Bedi GN, Pradeep U, Gupta A, Kasat PR, Sachani P. Revolutionizing reproduction: the impact of robotics and artificial intelligence (AI) in assisted reproductive technology: a comprehensive review. *Cureus* 2024;16(6):e63072. [\[CrossRef\]](#)
5. Kakkar P, Gupta S, Paschopoulos KI, Paschopoulos I, Siafaka V, et al. The integration of artificial intelligence in assisted reproduction: a comprehensive review. *Front Reprod Health* 2025;7:1520919. [\[CrossRef\]](#)
6. Zhang Q, Liang X, Chen Z. A review of artificial intelligence applications in *in vitro* fertilization. *J Assist Reprod Genet* 2025;42(1):3-14. [\[CrossRef\]](#)
7. Palmer GA, Kratka C, Szvetcz S, Fiser G, Fiser S, Sanders C, et al. Comparison of 36 assisted reproduction laboratories monitoring environmental conditions and instrument parameters using the same quality-control application. *Reprod Biomed Online* 2019;39(1):63-74. [\[CrossRef\]](#)
8. Wu YC, Chia-Yu Su E, Hou JH, Lin CJ, Lin KB, Chen CH. Artificial intelligence and assisted reproductive technology: A comprehensive systematic review. *Taiwan J Obstet Gynecol* 2025;64(1):11-26. [\[CrossRef\]](#)
9. Bormann CL, Thirumalaraju P, Kanakasabapathy MK, Kandula H, Souter I, Dimitriadis I, et al. Consistency and objectivity of automated embryo assessments using deep neural networks. *Fertil Steril* 2020;113(4):781-7.e1. [\[CrossRef\]](#)
10. Orovou E, Tzamourta KD, Tzitoridou-Chatzopoulou M, Kakatosi A, Sarantaki A. Artificial Intelligence in Assisted Reproductive Technology: A New Era in Fertility Treatment. *Cureus* 2025;17(4):e81568. [\[CrossRef\]](#)
11. Correa N, Cerquides J, Arcos JL, Vassena R. Supporting first FSH dosage for ovarian stimulation with machine learning. *Reprod Biomed Online* 2022;45(5):1039-45. [\[CrossRef\]](#)
12. Fanton M, Nutting V, Rothman A, Maeder-York P, Hariton E, Barash O, et al. An interpretable machine learning model for individualized gonadotrophin starting dose selection during ovarian stimulation. *Reprod Biomed Online* 2022;45(6):1152-9. [\[CrossRef\]](#)
13. Mathur P, Kakwani K, Diplav, Kudavelly S, Rama Raju GA. Deep Learning based quantification of ovary and follicles using 3D transvaginal ultrasound in assisted reproduction. *Annu Int Conf IEEE Eng Med Biol Soc.* 2020;2020:2109-12. [\[CrossRef\]](#)

14. Li H, Fang J, Liu S, Liang X, Yang X, Mai Z, et al. CR-Unet: A composite network for ovary and follicle segmentation in ultrasound images. *IEEE J Biomed Health Inform* 2020;24(4):974-83. [\[CrossRef\]](#)
15. Targosz A, Przystałka P, Wiaderkiewicz R, Mrugacz G. Semantic segmentation of human oocyte images using deep neural networks. *Biomed Eng Online* 2021;20(1):40. [\[CrossRef\]](#)
16. Firuzinia S, Afzali SM, Ghasemian F, Mirroshandel SA. A robust deep learning-based multiclass segmentation method for analyzing human metaphase II oocyte images. *Comput Methods Programs Biomed* 2021;201:105946. [\[CrossRef\]](#)
17. Hariton E, Chi EA, Chi G, Morris JR, Braatz J, Rajpurkar P, et al. A machine learning algorithm can optimize the day of trigger to improve *in vitro* fertilization outcomes. *Fertil Steril* 2021;116(5):1227-35. [\[CrossRef\]](#)
18. Letterie G, MacDonald A, Shi Z. An artificial intelligence platform to optimize workflow during ovarian stimulation and IVF: process improvement and outcome-based predictions. *Reprod Biomed Online* 2022;44(2):254-60. [\[CrossRef\]](#)
19. Nixon B, Schjenken JE, Burke ND, Skerrett-Byrne DA, Hart HM, De luliis GN, et al. New horizons in human sperm selection for assisted reproduction. *Front Endocrinol (Lausanne)* 2023;14:1145533. [\[CrossRef\]](#)
20. Gatimel N, Moreau J, Parinaud J, Léandri RD. Sperm morphology: assessment, pathophysiology, clinical relevance, and state of the art in 2017. *Andrology* 2017;5(5):845-62. [\[CrossRef\]](#)
21. Ottl S, Amiriparian S, Gerczuk M, Schuller BW. motilitAI: A machine learning framework for automatic prediction of human sperm motility. *iScience* 2022;25(8):104644. [\[CrossRef\]](#)
22. Riordon J, McCallum C, Sinton D. Deep learning for the classification of human sperm. *Comput Biol Med* 2019;111:103342. [\[CrossRef\]](#)
23. Sato T, Kishi H, Murakata S, Hayashi Y, Hattori T, Nakazawa S, et al. A new deep-learning model using YOLOv3 to support sperm selection during intracytoplasmic sperm injection procedure. *Reprod Med Biol* 2022;21(1):e12454. [\[CrossRef\]](#)
24. Mendizabal-Ruiz G, Chavez-Badiola A, Aguilar Figueroa I, Martinez Nuño V, Flores-Saiffe Farias A, Valencia-Murilloa R, et al. Computer software (SiD) assisted real-time single sperm selection associated with fertilization and blastocyst formation. *Reprod Biomed Online* 2022;45(4):703-11. [\[CrossRef\]](#)
25. Cissen M, Wely MV, Scholten I, Mansell S, Bruin JP, Mol BW, et al. Measuring sperm DNA fragmentation and clinical outcomes of medically assisted reproduction: a systematic review and meta-analysis. *PLoS One* 2016;11(11):e0165125. [\[CrossRef\]](#)
26. McCallum C, Riordon J, Wang Y, Kong T, You JB, Sanner S, et al. Deep learning-based selection of human sperm with high DNA integrity. *Commun Biol* 2019;2:250. [\[CrossRef\]](#)
27. Dimitriadis I, Bormann CL, Kanakasabapathy MK, Shafiee H, Thirumalaraju P, Gupta R, et al. Deep convolutional neural networks (CNN) for assessment and selection of normally fertilized human embryos at the pronuclear stage. *Fertil Steril* 2019;112:e272. [\[CrossRef\]](#)
28. Jiang VS, Bormann CL, Hariton E, Pavlovic ZJ, Fanton M, Chi EA, et al. Artificial intelligence in the *in vitro* fertilization laboratory: a review of advancements over the last decade. *Fertil Steril* 2023;120(1):17-23. [\[CrossRef\]](#)
29. Bulletti C, Franasiak JM, Busnelli A, Sciorio R, Berrettini M, Aghajanova L, et al. Artificial intelligence, clinical decision support algorithms, mathematical models, calculators applications in infertility: systematic review and hands on digital applications. *Mayo Clin Proc Digit Health*. 2024;2(4):518-32. [\[CrossRef\]](#)
30. Cimadomo D, Marconetto A, Trio S, Chiappetta V, Innocenti F, Albricci L, et al. Human blastocyst spontaneous collapse is associated with worse morphological quality and higher degeneration and aneuploidy rates: a comprehensive analysis standardized through artificial intelligence. *Hum Reprod* 2022;37(10):2291-2306. [\[CrossRef\]](#)
31. Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, et al. Deep learning enables robust assessment and selection of human blastocysts after *in vitro* fertilization. *NPJ Digit Med* 2019;2:21. [\[CrossRef\]](#)
32. Hodes-Wertz B, Grifo J, Ghadir S, Kaplan B, Laskin CA, Glassner M, et al. Idiopathic recurrent miscarriage is caused mostly by aneuploid embryos. *Fertil Steril* 2012;98(3):675-80. [\[CrossRef\]](#)
33. Kato K, Ueno S, Berntsen J, Kragh MF, Okimura T, Kuroda T. Does embryo categorization by existing artificial intelligence, morphokinetic or morphological embryo selection models correlate with blastocyst euploidy rates? *Reprod Biomed Online* 2023;46(2):274-81. [\[CrossRef\]](#)
34. Jiang VS, Kandula H, Thirumalaraju P, Kanakasabapathy MK, Cherouveim P, Souter I, et al. The use of voting ensembles to improve the accuracy of deep neural networks as a non invasive method to predict embryo ploidy status. *J Assist Reprod Genet* 2023;40(2):301-8. [\[CrossRef\]](#)
35. Pirtea P, Scott RT Jr, de Ziegler D, Santibañez J, Bormann CL, Dimitriadis I, et al. Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF. *Hum Reprod*. 2022;37(8):1746-59. [\[CrossRef\]](#)
36. Fitz VW, Kanakasabapathy MK, Thirumalaraju P, Kandula H, Ramirez LB, Boehnlein L, et al. Should there be an "AI" in TEAM? Embryologists selection of high implantation potential embryos improves with the aid of an artificial intelligence algorithm. *J Assist Reprod Genet* 2021;38(10):2663-70. [\[CrossRef\]](#)
37. Goyal A, Kuchana M, Ayyagari KPR. Machine learning predicts live-birth occurrence before in-vitro fertilization treatment. *Sci Rep* 2020;10(1):20925. [\[CrossRef\]](#)
38. Mendizabal Ruiz G, Chavez Badiola A, Mendizabal Ruiz A, Aguilar Figueroa I, Martinez Nuño V, Flores Saiffe Farias A, et al. A digitally controlled, remotely operated ICSI system: case report of the first live birth. *Reprod Biomed Online* 2025;50(4):607-11. [\[CrossRef\]](#)



# Artificial Intelligence–Driven Weight Management: Current Evidence and Clinical Implications

 Nursena Nahya Servi

Department of Nutrition and Dietetics, Faculty of Health Sciences, Lokman Hekim University, Ankara, Türkiye

## Abstract

Obesity remains one of the most urgent global public health challenges, necessitating innovative and scalable strategies for effective weight management. This narrative review aims to synthesize current evidence (2021–2026) on the role of artificial intelligence (AI) in weight loss and obesity management, and to evaluate its clinical potential, limitations, and future directions. Recent advances in AI, including machine learning and deep learning techniques, have introduced novel opportunities for personalized nutrition, predictive modeling, and digitally supported behavioral interventions. The literature indicates that AI-driven systems show substantial potential in predictive weight loss modeling, reinforcement learning–based treatment optimization, digital coaching platforms, and biomarker-integrated personalization strategies. Importantly, while AI technologies may enhance scalability and personalization, they should be positioned as clinical decision-support tools rather than replacements for dietitians and healthcare professionals. However, the field remains heterogeneous, with a limited number of long-term randomized controlled trials, variable methodological transparency, and insufficient external validation of predictive models. While AI technologies may enhance scalability and personalization, they should be positioned as clinical decision-support tools rather than replacements for dietitians and healthcare professionals. Ethical considerations, data governance, and algorithmic transparency remain critical for safe and responsible implementation. Overall, AI represents a promising adjunct in weight management; however, its integration into clinical nutrition practice requires rigorous validation and interdisciplinary collaboration.

**Keywords:** Artificial intelligence; Digital health; Machine learning; Obesity management; Weight loss

Modern lifestyle trends, including physical inactivity and unhealthy dietary patterns, have been strongly associated with obesity, a major health condition that substantially increases the risk of multiple pathologies.<sup>[1]</sup> Today, obesity represents one of the most critical global pandemics facing society, with prevalence rates reaching alarming and unacceptable levels worldwide.<sup>[2]</sup> The seriousness of this issue was explicitly recognized at the 75<sup>th</sup> World Health Assembly in 2022, where Member States

endorsed new recommendations for obesity prevention and management and approved the World Health Organization's Acceleration Plan to Stop Obesity. Without effective action against obesity, achieving the target of reducing premature mortality from non-communicable diseases (NCDs) by 30% by 2030 appears unlikely.<sup>[3]</sup> These developments clearly demonstrate that obesity is not merely an individual health concern but a deepening global public health crisis that demands innovative and systemic solutions.<sup>[4]</sup>

**Cite this article as:** Servi NN. Artificial Intelligence–Driven Weight Management: Current Evidence and Clinical Implications. *Lokman Hekim Health Sci* 2026;6(2):360–371.

**Correspondence:** Nursena Nahya Servi, PhD. Lokman Hekim Üniversitesi, Sağlık Bilimleri Fakültesi, Beslenme ve Diyetetik Anabilim Dalı, Ankara, Türkiye  
**E-mail:** nursenanahya@gmail.com **Submitted:** 05.03.2026 **Revised:** 19.04.2026 **Accepted:** 30.04.2026 **Available Online:** 21.05.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



Projections from the World Obesity Atlas 2025 further underscore the magnitude of the problem. By 2030, approximately 50% of adult men and women are expected to have an elevated body mass index (BMI). Obesity prevalence is projected to reach 17% among men and 22% among women. Absolute number projections suggest that more than 2.9 billion adults worldwide will be living with high BMI by 2030, including approximately 1.1 billion individuals with obesity (BMI  $\geq 30$  kg/m<sup>2</sup>), of whom an estimated 487 million will be men and 643 million women. Notably, the number of individuals with class II obesity or higher (BMI  $\geq 35$  kg/m<sup>2</sup>) is expected to approach 400 million, with women constituting the majority of this group. These projections are critically important for healthcare planning and resource allocation, particularly given that two-thirds of adults with BMI  $\geq 35$  kg/m<sup>2</sup> are anticipated to reside in low- and middle-income countries.<sup>[5]</sup> This trend indicates that the demand for obesity-related clinical interventions will intensify globally, disproportionately burdening resource-constrained health systems.

In Türkiye, projections indicate that by 2025, 71% of the adult population will have elevated BMI ( $\geq 25$  kg/m<sup>2</sup>), and 36% will be living with obesity. If current trends continue, the number of adults with high BMI is expected to reach approximately 47.4 million by 2030. A substantial proportion of premature deaths and disability-adjusted life years attributable to high BMI are linked to Type 2 diabetes, cardiovascular diseases, and various cancers, highlighting obesity's central role in the burden of NCDs.<sup>[5]</sup> Although the state plays an important role in combating obesity in Türkiye, its failure to fully adopt the advanced health policies and long-term strategic approaches seen in Europe has led to structural deficiencies in effectively tackling obesity.<sup>[6]</sup> Collectively, these data demonstrate that obesity cannot be explained solely by individual choices; rather, health-related behaviors must be addressed within a multidimensional and systemic framework.

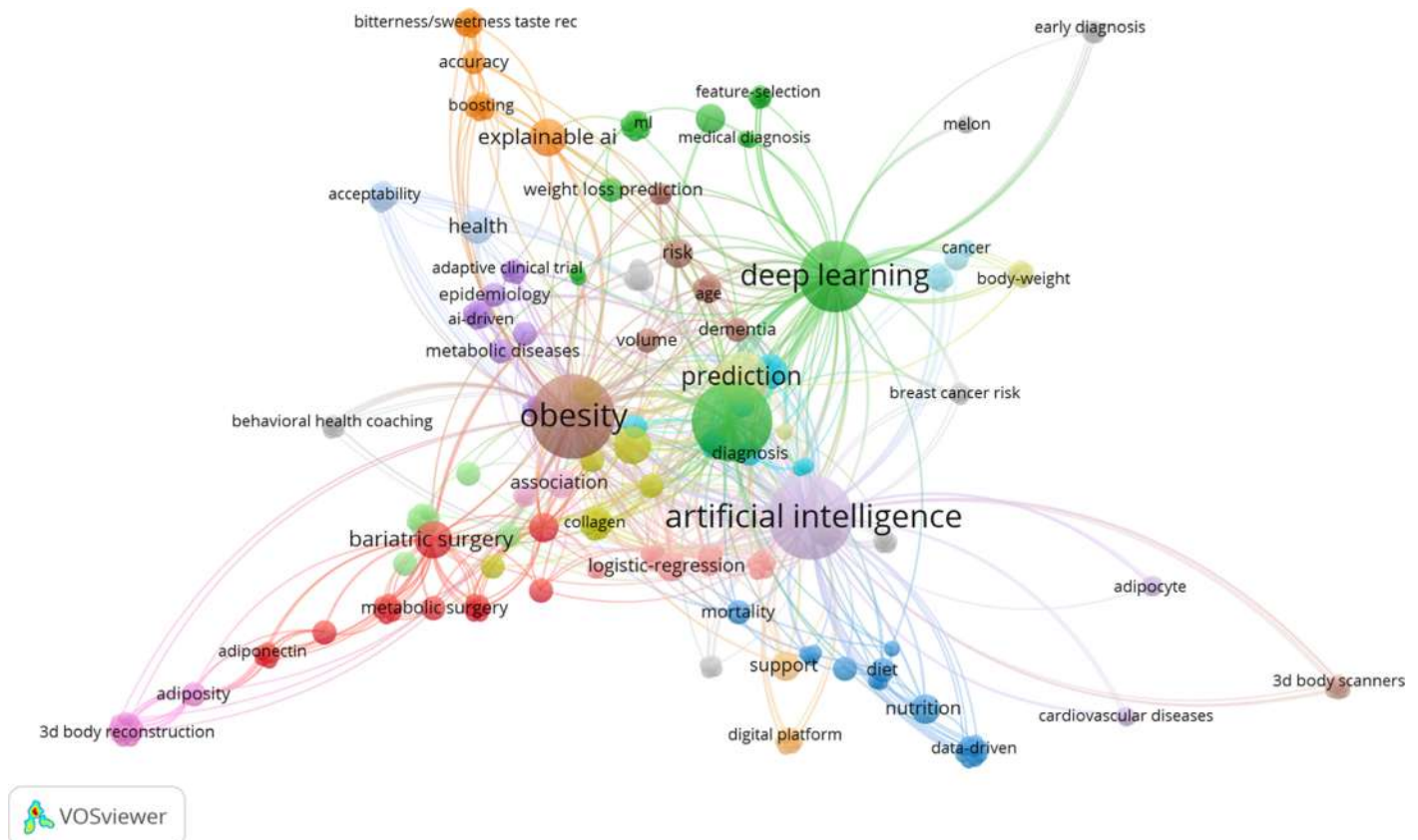
Health-related behaviors play a fundamental role in improving population health globally. Adherence to healthy dietary patterns, sufficient physical activity, adequate sleep, and avoidance of risk behaviors such as smoking significantly reduce the risk of chronic diseases and all-cause mortality, while also exerting positive effects on mental health. Nevertheless, unhealthy behaviors remain highly prevalent and continue to impose a substantial burden on healthcare systems. In this context, dietitians – who provide evidence-based approaches to the regulation of nutritional behaviors – occupy a pivotal role in supporting sustainable health behavior change.<sup>[7]</sup>

Nutrition and dietetics constitute a critical discipline in the prevention and management of obesity and obesity-related chronic diseases in modern societies. Irregular eating habits driven by demanding work environments, increased consumption of ultra-processed foods, and sedentary lifestyles have further amplified the importance of dietitians in promoting sustainable healthy living.<sup>[8–10]</sup> Beyond disease prevention, nutrition is increasingly recognized as a cornerstone of treatment, with the “food as medicine” paradigm suggesting that dietary interventions may offer safer and more sustainable solutions in obesity management compared with purely pharmacological or surgical approaches.<sup>[11,12]</sup>

Within this framework, dietitians are positioned as essential primary healthcare providers capable of delivering equitable, respectful, and effective care to individuals with overweight or obesity. Strengthening dietitians' education and professional competencies is crucial to implementing person-centered approaches that account for the biological, behavioral, and environmental determinants contributing to excess weight.<sup>[11]</sup>

Despite the growing global burden of obesity, traditional weight management approaches based solely on conventional dietitian counseling may exhibit limited long-term effectiveness in certain contexts. Standardized “one-size-fits-all” nutritional interventions often fail to adequately capture interindividual biological, behavioral, and environmental variability, thereby limiting sustainable outcomes. This reality underscores the need to support dietitian-led services with more personalized, dynamic, and continuously monitored models of care.<sup>[8,13,14]</sup>

In recent years, rapid advances in digital health technologies have positioned artificial intelligence (AI) as a promising complementary tool in dietetic practice. AI-based systems have demonstrated potential in generating individualized nutrition recommendations, supporting behavior change, monitoring physical activity, and predicting weight loss success.<sup>[15]</sup> However, despite its promising role in combating obesity, significant gaps remain in the literature regarding its practical implementation. The majority of studies have focused primarily on obesity detection, whereas prevention and long-term management strategies remain underexplored.<sup>[16]</sup> By leveraging the capacity to analyze large-scale datasets and provide individualized feedback, AI-based approaches may contribute to the development of more effective, sustainable, and scalable dietitian-guided obesity management strategies.<sup>[15]</sup>



**Figure 1.** Keyword co-occurrence network based on Web of Science–indexed publications addressing artificial intelligence applications in obesity and weight loss. The network visualizes conceptual relationships among keywords extracted from the literature. Node size represents keyword frequency, link thickness reflects the strength of co-occurrence, and colors indicate major thematic clusters within the research field.

The aim of this narrative review is to critically evaluate current evidence on AI –driven weight management, with a focus on its clinical applicability, predictive capabilities, and integration into dietetic practice.

## Materials and Methods

### Review Design

This study was conducted as a narrative review aimed at providing a comprehensive and critical interpretation of recent scientific evidence on the role of AI in weight loss, obesity management, and related lifestyle interventions. Rather than applying a systematic review framework, this study adopts an interpretative approach to synthesize current knowledge, identify emerging trends, and highlight key strengths and limitations within the existing literature.

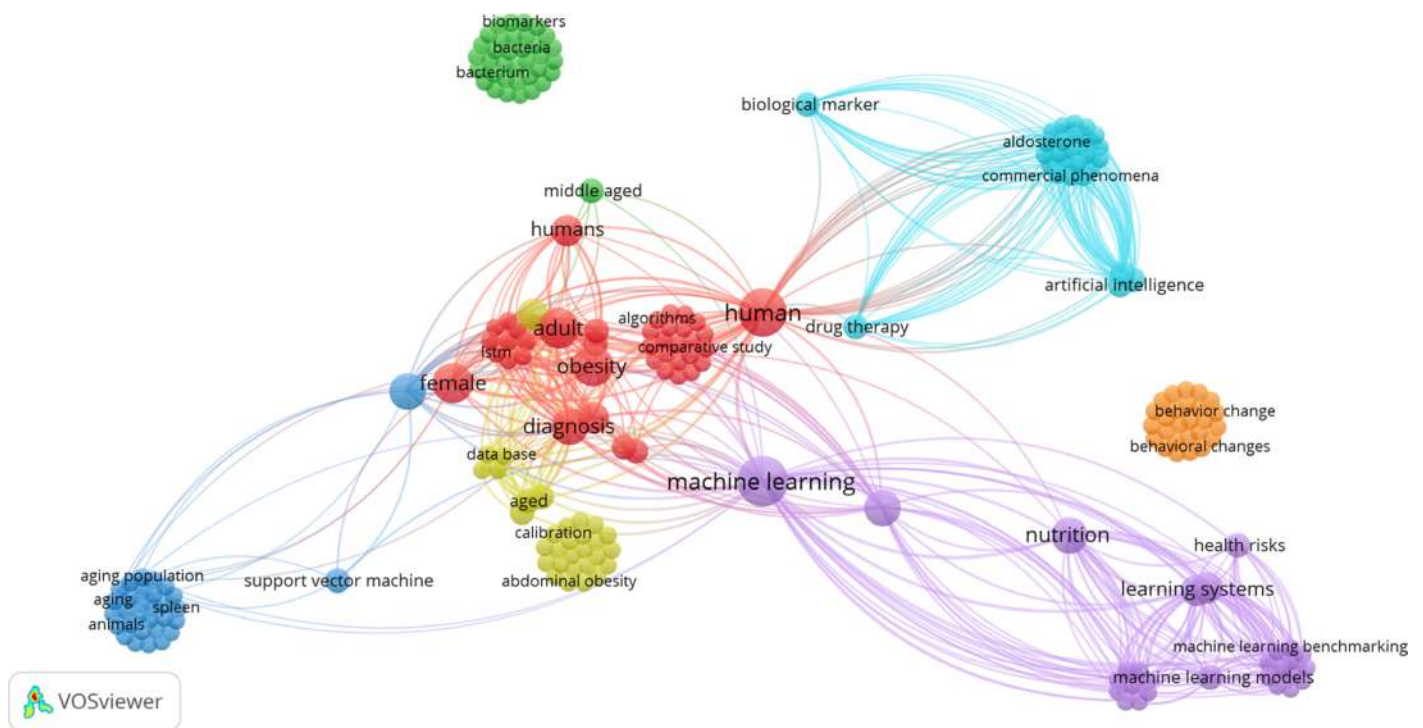
### Search Strategy

Relevant literature was identified through searches in the Web of Science and Scopus databases using the keywords (“artificial intelligence” OR “machine learning” OR “deep learning” OR “AI”) AND (“weight loss” OR obesity OR “body

weight” OR “weight management”). The search was limited to articles published in English between 2021 and 2026 to ensure coverage of the most recent developments in AI technologies and their clinical relevance. The literature search was conducted up to April 2026. Studies were selected based on their relevance to the topic, with priority given to research providing empirical findings or substantial conceptual contributions to AI-driven weight management. Both clinical and experimental studies were considered. Unlike applying strict inclusion–exclusion criteria as in systematic reviews, the selection process was guided by the aim of capturing representative and informative studies reflecting the diversity of current research in the field. The findings were synthesized using a thematic approach.

### Bibliometric Analysis

To support the methodological framework of this review, a term co-occurrence network was constructed based on the current scientific literature in the field of AI and weight loss. VOSviewer software (version 1.6.20) was used to visualize conceptual relationships and thematic clusters among keywords.



**Figure 2.** Keyword co-occurrence network based on Scopus-indexed publications examining artificial intelligence–related approaches in obesity and weight loss research. The network illustrates conceptual relationships among frequently co-occurring keywords. Node size represents keyword frequency, link thickness indicates the strength of co-occurrence, and colors denote distinct thematic clusters within the literature.

The visualization presents thematic clusters based on co-occurrence strength, where node size represents term frequency and link thickness indicates the strength of the association between terms. The resulting network structure illustrates how the concepts of AI, machine learning (ML), and deep learning (DL) are positioned within specific thematic contexts in predictive approaches to weight loss in the existing literature.

The dataset used for visualization was retrieved from the Web of Science (Fig. 1) and Scopus (Fig. 2) databases using the expanded search query: (“artificial intelligence” OR “machine learning” OR “deep learning” OR “AI”) AND (“weight loss” OR obesity OR “body weight” OR “weight management”).

A total of 20,636 documents were retrieved from Web of Science and 8,295 from Scopus (2021–2026, English). Term co-occurrence analysis was performed using a minimum occurrence threshold of 5 and the full counting method. Only terms meeting co-occurrence and connectivity criteria were included in the final visualization. Consequently, 289 items from Web of Science and 279 items from Scopus were retained and visualized as part of the largest connected network.

This bibliometric analysis was used as a complementary exploratory tool to provide insight into the conceptual

structure and methodological trends in AI-based weight management research.

Figure 1 presents a Web of Science–based keyword co-occurrence network highlighting the conceptual structure of AI-driven research in obesity and weight loss. The central positioning of AI, obesity, and prediction indicates a strong emphasis on predictive modeling approaches, particularly DL –based methods. Notably, the presence of explainable AI alongside accuracy- and feature selection–related terms reflects a growing interest in model interpretability and clinical transparency. Connections with nutrition, diagnosis, and bariatric surgery further demonstrate the integration of AI applications into both clinical and lifestyle-oriented weight management contexts.

Figure 2 illustrates a Scopus-based keyword co-occurrence network that further refines the conceptual landscape of AI –driven research in obesity and weight loss. In this network, human- and obesity-related terms occupy a central position, closely connected with ML and algorithm-based approaches, indicating a strong focus on human-centered predictive modeling. Distinct clusters related to nutrition and learning systems, biological markers and hormonal pathways, and behavioral change highlight the multidimensional nature of current research. Notably,

the separation of clusters associated with biomarkers and behavior change suggests that biological and behavioral components of weight management are often addressed in parallel rather than within fully integrated AI frameworks.

## AI and Weight Loss

### Conceptual Framework of AI in Nutrition

Although the terms AI, ML, and DL are often used interchangeably today, there is actually a hierarchical relationship between these concepts. AI refers to machines that are programmed to perform tasks that normally require human intelligence, such as pattern recognition, learning from experience, decision-making, and problem-solving, and that can exhibit intelligent behavior. ML is a subfield of AI that enables machines to learn from data without being explicitly programmed and to independently improve their performance over time. DL, as a subfield of ML, focuses on multi-layered artificial neural networks and enables the identification of complex patterns within large data sets. While DL has contributed to significant advances in application areas such as image and speech recognition and autonomous vehicles, interpreting and explaining how these models work remains a major challenge.<sup>[17]</sup> Beyond the conceptual hierarchy of AI technologies, another critical dimension is the degree of system autonomy in decision-making processes.

Mennella et al.<sup>[18]</sup> define the autonomy levels of AI systems in five stages from 0 to 4. Level 0 indicates the absence of AI, and standard care is performed entirely by humans; the final decision-maker is human. At Level 1, AI provides recommendations to the system, but the final decision regarding treatment and therapy is made by the clinician. At Level 2, AI generates clinical decisions; however, the process is continuously monitored by humans, and the ultimate responsibility remains with humans. At Level 3, the system makes clinical decisions autonomously, without continuous human supervision; however, human support may be involved in cases of uncertainty. At this level, operational decisions are made by AI, but a human backup mechanism is available. Level 4 represents the highest level of autonomy; AI makes clinical decisions completely independently, and there is no human backup system. At this stage, the final decision-maker is directly AI. In the context of healthcare applications, particularly nutrition recommendation systems, these technological and autonomy distinctions become highly relevant.

Nutrition recommendation systems are classified into three main categories: Knowledge-based, AI-based, and

large language model (LLM)-based. Knowledge-based systems offer reliability, transparency, and strong data privacy by relying on established nutrition guidelines; AI-based systems can uncover hidden patterns by processing large volumes of data. LLM-based systems, on the other hand, have the potential to provide personalized nutrition recommendations thanks to their ability to leverage extensive internet resources and interactive communication capabilities.<sup>[19]</sup>

### AI-Supported Interventions and Predictive Personalization in Weight Management

AI-supported personalized nutrition is expected to be one of the greatest revolutions in both nutrition and health. A feasibility study compared manual (questionnaire-based) and wearable sensor-based automatic data collection methods within the scope of a structured weight loss-focused nutritional intervention applied to overweight and obese individuals. At the end of the intervention, participants showed an average weight loss of 2 kg, along with significant improvements in BMI, visceral fat, waist circumference, total cholesterol, and hemoglobin A1c levels. The study demonstrated that data obtained from wearable devices provides a viable and reliable infrastructure for the development and validation of AI-based clinical nutrition applications.<sup>[20]</sup> While behavioral weight loss, considered the gold standard, achieves clinically meaningful weight loss (7–10%), its high costs and limited scalability make it difficult to reach large populations. It has been demonstrated that a reinforcement learning (RL)-based AI system can allocate limited resources in the most effective and cost-efficient manner by adjusting the intensity of intervention in an individualized and time-sensitive manner. Unlike traditional stepwise care models, RL algorithms can continuously and automatically re-optimize treatment and consider both effectiveness and cost in intervention selection. The findings indicate that this AI-supported approach has the potential to deliver weight loss outcomes comparable to gold standard treatments at a lower cost and to increase the scalability of obesity treatment.<sup>[21]</sup>

Digital coaching studies show that AI-powered systems, especially when integrated into hybrid human-AI frameworks, can increase both scalability and personalization in weight management. On the online weight loss coaching platform, AI-based optimization methods, including automated food diary feedback and ML-generated meal planning, have been shown to reduce the coach's workload while maintaining the effectiveness of personalized guidance.<sup>[4]</sup> Importantly, the study did

not eliminate human coaching but rather repositioned AI as a layer that enhances efficiency in the interaction process. This demonstrates that algorithmic automation can support operational scalability without compromising the quality of behavioral support. Complementing these findings, generational analyses of AI health coach interactions have revealed nuanced participation patterns across age groups. Although high interaction levels were observed across all generations, participation increased with age, with Boomers (ages 60–78) showing the highest participation and longest interaction duration. Younger users (Generation Z and Generation Y) were more inclined to initiate conversations spontaneously, while older groups predominantly participated in structured, guidance-based interactions. Notably, food diary feedback emerged as the most frequently discussed topic across all generations, reinforcing the central role of diet self-monitoring in AI-supported coaching. These findings emphasize that not only the frequency of interaction but also the style of interaction can influence the effectiveness of AI-focused interventions.<sup>[22]</sup> Beyond individual platform-based interventions, broader evidence from a comprehensive review further contextualizes the role of text-based AI chatbots in health behavior change. In this review, chatbots were most frequently used as routine coaches (62.8%) or on-demand assistants (27.9%) and generally relied on established behavior change techniques such as cognitive behavioral therapy and goal setting, feedback, monitoring, and social support. Notably, over half of the systems (53.5%) were developed using ready-made, no-code platforms, highlighting the increasing accessibility of chatbot development. Although positive effects were reported in 81.7% of comparisons, only 35.8% showed moderate or large effect sizes, and the evidence remained heterogeneous. Furthermore, while studies primarily focused on physical activity, diet, stress management, and smoking cessation, outcomes related to direct weight management and sleep were relatively underrepresented.<sup>[7]</sup> In this context, the AI-powered eTRIP application developed for overweight and obese individuals of Southeast Asian origin takes the focus on behavioral self-regulation a step further. The multi-component structure, which combines chatbot-based food trigger queries, food logging using an image recognition system adapted to local foods, and time-based reminders, has led to a significant reduction in overeating and snacking behaviors in a short period of time. The increase in physical activity levels and decrease in depression scores after the intervention suggest that digital self-monitoring tools can affect not only eating behavior

but also related psychosocial outcomes. However, the lack of a significant change in anxiety levels and the short duration of the intervention indicate that such applications should be carefully evaluated in terms of long-term clinical outcomes. The low dropout rate (8.4%) and positive user feedback support the system's feasibility.<sup>[23]</sup>

AI has been applied to weight management through both unstructured real-world data analysis and structured predictive modeling. In the off-label use of semaglutide, AI and natural language processing (NLP)-based analyses of user-generated comments showed that 31.2% reported weight loss, with an average reduction of approximately 26 kg over 3.5 months. Common side effects reported included nausea, vomiting, headache, fatigue, and dizziness, while topic modeling highlighted themes such as dietary changes, cost concerns, and long-term expectations.<sup>[24]</sup> While these findings provide valuable patient-centered information, relying on self-reported data online leads to potential reporting bias and limits causal interpretation. In contrast, AI models trained on structured digital health records offer a more controlled prediction framework. An XGBoost model built on 121,564 dietary entries from 1,088 individuals with type 2 diabetes predicted  $\geq 3\%$  weight loss with 93% accuracy. Shapley analysis highlighted the role of meal timing, suggesting that AI-supported nutritional analysis could advance diet personalization beyond simple nutrient quantities.<sup>[25]</sup> However, external validation and cross-population reproducibility are essential for transitioning to clinical practice. Together, these studies demonstrate that AI can both extract experiential signals from real-world patient narratives and generate predictive, personalized dietary strategies. When interpreted in clinically controlled contexts, these complementary approaches can increase precision in weight management.

Predictive modeling approaches clearly demonstrate how AI can develop personalized weight management strategies. In a medical weight management program, baseline variables capable of predicting clinically meaningful weight loss ( $>10\%$ ) over 12 months were identified using ML methods. The random forest model identified quality of life (SF-36 mental and physical scores), age, socioeconomic status, and excess body weight as the strongest predictors of weight loss success.<sup>[26]</sup> These findings demonstrate that weight loss is strongly linked not only to biometric indicators but also to psychosocial and functional factors. Therefore, considering multidimensional data in initial assessments could provide a significant advantage in planning AI-based personalized interventions.

When transitioning from predictive models to intervention, the clinical outcomes of AI-guided personalized diet support prescriptions were also examined. In a randomized controlled pilot study, the average weight loss at 180 days was 12.3% in the AI group, while it was 7.2% in the physician-guided standard approach, revealing a significant difference of 5.1% between the groups.<sup>[27]</sup> The rate of clinically significant weight loss ( $\geq 5\%$ ) reached 83.3% in the AI group, and this approach was associated with a 4.4 times higher probability of success. In addition, more pronounced improvements were observed in BMI, fat mass, and visceral fat indicators. Although both groups followed the same diet and exercise protocol, the holistic integration of genetic, metabolic, and behavioral data through AI appears to have provided an additional benefit. However, due to the pilot design and limited sample size, the results need to be validated in larger populations.

The role of AI in weight management is not limited to individual prediction and intervention; it also extends to the production and optimization processes of nutritional components. A review examining AI-supported bioprocesses for obtaining concentrated nutritional components from different biological sources reports that ML models can increase yield, stability, and bioavailability by predicting process parameters and providing real-time control.<sup>[13]</sup> This approach is considered important in bridging the gap between biological potential and personalized nutrition applications. However, issues such as regulation, model transparency, scalability, and consumer acceptance remain key obstacles to be overcome in the transition to clinical application.

Similarly, AI-supported medical devices are also emerging as complementary tools for long-term weight management. The ELECT study reported that the use of a non-pharmacological, Food and Drug Administration-approved super-absorbent hydrogel capsule in combination with lifestyle counseling resulted in an average of 11.2% sustained weight loss over 48 weeks.<sup>[28]</sup> The high compliance rate ( $>94\%$ ) and the absence of serious device-related adverse events suggest that such devices may be safe and tolerable. However, since it is difficult to separate the independent effect of the device from lifestyle counseling, the results must be evaluated within the context of a comprehensive treatment framework.

Studies aimed at capturing the biopsychosocial nature of weight loss at the algorithmic level are also noteworthy. Including sleep quality and stress levels in the artificial neural network model has significantly increased the

accuracy of weight loss predictions.<sup>[2]</sup> When these variables are removed from the model, the significant increase in error rates supports the effect of psychological and behavioral factors on metabolic outcomes.

AI has been increasingly utilized in obesity research for measurement, prediction, and intervention, with numerous studies demonstrating its superior ability to identify clinically meaningful patterns and achieve higher predictive accuracy compared to traditional statistical approaches.<sup>[29]</sup>

A recent systematic review demonstrated that most AI-based studies focus on identifying obesity risk with high accuracy, while real-world implementation, prevention strategies, and clinical integration remain relatively underexplored compared to predictive applications.<sup>[30]</sup> This gap highlights a critical translational limitation, as high-performing models do not necessarily translate into effective, scalable interventions in clinical practice. In parallel, emerging evidence indicates that the true transformative potential of AI lies in its ability to integrate multi-omics data, enabling a systems-level understanding of obesity beyond traditional anthropometric measures. AI-driven multi-omics approaches have been shown to improve predictive performance by 5–15% and uncover complex biological interactions across genomic, metabolic, and environmental layers, thereby facilitating more precise risk stratification and personalized intervention strategies.<sup>[31]</sup> Nevertheless, substantial challenges remain, including data heterogeneity, lack of standardization, limited external validation, and the “black-box” nature of many models, all of which hinder clinical translation. Taken together, these findings suggest that while AI holds considerable promise in obesity management, its future impact will depend on moving beyond isolated predictive models toward integrative, clinically applicable frameworks that combine multi-dimensional data with real-world decision-making processes.

In this context, emerging approaches in the current literature provide concrete examples of how this transformation can be achieved. Specifically, one line of research focuses on the AI-driven identification of neurobiological signatures of obesity and weight loss, while another emphasizes the real-time monitoring of behavioral patterns through internet of things (IoT)-integrated systems.<sup>[32–34]</sup> Recent evidence further supports the emerging role of AI in uncovering the neurobiological underpinnings of obesity. In a DL-based neuroimaging study, BMI was successfully inferred from structural brain magnetic resonance imaging

(MRI), and importantly, AI-predicted reductions in BMI were significantly associated with actual weight loss following an 18-month lifestyle intervention. Notably, these neural BMI changes were more pronounced in Mediterranean diet-based intervention groups, suggesting that dietary patterns may exert measurable effects on brain structure beyond traditional metabolic outcomes. Moreover, AI-derived BMI estimates were independently associated with key metabolic risk markers, including visceral adiposity, liver fat, insulin resistance, and high-density lipoprotein-C levels, even after adjusting for observed BMI.<sup>[33]</sup> In parallel, a recent DL study has demonstrated that convolutional neural networks applied to structural MRI data can predict BMI with substantially higher accuracy than traditional ML approaches, suggesting.<sup>[34]</sup> These findings highlight that AI-driven approaches may capture a broader, neurobiologically embedded phenotype of obesity rather than relying solely on anthropometric measures. However, the relatively modest strength of the association between predicted and actual weight loss indicates that such models should currently be interpreted as complementary tools rather than standalone clinical predictors. In addition, an IoT-integrated framework demonstrated that wearable sensors combined with ML algorithms can accurately classify daily physical activities with high precision (~98%), enabling continuous tracking of physical activity patterns and dietary behaviors. Such systems allow real-time integration of energy intake and expenditure, offering a dynamic perspective on weight management that extends beyond static clinical assessments. Particularly in post-bariatric surgery populations, this approach may provide critical support for long-term adherence to dietary and physical activity recommendations.<sup>[32]</sup>

From a clinical dietetics perspective, the integration of AI into practice is particularly relevant for patient populations requiring continuous monitoring and individualized intervention, such as those with obesity, Type 2 diabetes, and other metabolic disorders.<sup>[35-37]</sup> Recent evidence demonstrates that AI-driven systems, including ML- and DL-based tools, enable real-time monitoring, predictive modeling, and personalized treatment planning, thereby improving disease management and patient outcomes. For instance, in diabetes care, AI-supported technologies such as continuous glucose monitoring systems, smart insulin delivery devices, and decision-support algorithms facilitate dynamic adjustment of dietary and pharmacological interventions based on real-time metabolic data.<sup>[37]</sup>

Beyond diabetes, AI applications extend to broader metabolic disorders, including obesity and metabolic

dysfunction-associated conditions, where large-scale data integration from wearable devices, electronic health records, and lifestyle inputs allows for precise risk stratification and tailored nutritional recommendations. These systems can identify high-risk individuals, predict disease progression, and support early intervention strategies, which are critical for long-term dietary adherence and prevention of complications. Furthermore, AI-enhanced mobile applications and digital health platforms have been shown to improve glycemic control, promote weight loss, and support behavioral modification through continuous feedback mechanisms.<sup>[35,36]</sup>

Taken together, these findings suggest that the most immediate clinical application of AI in dietetic practice lies in high-risk populations requiring sustained lifestyle modification, where continuous monitoring and adaptive, data-driven interventions are essential.

Representative studies highlighting the diversity of AI applications in weight management are summarized in Table 1.

Finally, a recent study shows that the traditional and generalized “total restriction” approach to BMI reduction does not adequately reflect individual response differences. Personalized optimal dietary patterns developed using meta-algorithms have been found to be more effective in reducing BMI compared to general recommendations, highlighting the importance of heterogeneous response patterns in weight management. Notably, the fact that higher consumption of certain foods or beverages is associated with lower BMI in some subgroups suggests that dietary interventions cannot be explained by linear and unidirectional effect assumptions. ML models that predict heterogeneous treatment effects can more clearly distinguish between personalized and general approaches, suggesting that AI-based decision support systems may offer a potential advantage in capturing individual metabolic and behavioral differences.<sup>[14]</sup> However, such algorithmic outputs should not be evaluated independently of the clinical context, and caution should be exercised when drawing causal inferences. Therefore, AI-based meta-algorithms should not be positioned as tools that will replace dietitians; rather, they should be positioned as decision support systems used under the guidance of dietitians.

A conceptual framework proposed to systematically organize complex AI applications in the literature demonstrates that AI, ML, and DL techniques can be integrated into every stage of the nutrition process.

**Table 1.** Comparison of the means of lung function parameters of the different categories of Vitamin D level

Lacruz-Pleguezuelos et al. <sup>[20]</sup>	Randomized crossover controlled trial (feasibility study)	Overweight/obese adults (n=93)	ML-supported wearable integration (CGM, smartwatch, food imaging)	Monitoring	Significant improvements in body weight, metabolic parameters, and glycemic control; ML identified distinct glucose response patterns	Demonstrates the potential of AI-driven wearable systems for real-time monitoring and personalized nutrition strategies
Forman et al. <sup>[21]</sup>	Randomized controlled trial protocol	Adults with overweight/obesity (planned n=336)	Reinforcement learning (RL)	Decision-making	RL dynamically optimized intervention intensity, achieving comparable weight loss with reduced resource utilization	Demonstrates the potential of AI to optimize treatment allocation and improve cost-effectiveness in weight management
Chew et al. <sup>[40]</sup>	Single-group pre-test-post-test (mixed methods)	Adults with overweight/obesity (n=230)	Chatbot+computer vision+behavioral nudging	Behavioral intervention	Improvements in overeating, self-regulation, physical activity, and psychological outcomes with high engagement	Demonstrates the potential of AI-assisted behavioral tools for scalable and real-time lifestyle intervention
Pokushalov et al. <sup>[27]</sup>	Randomized controlled trial	Adults with overweight/obesity (n=60)	Multi-omics AI integration (genetic, metabolic, behavioral data)	Precision treatment	Greater weight loss and improved body composition compared to standard care; higher responder rates	Demonstrates the potential of AI-driven precision nutrition to enhance treatment effectiveness
Alsareii et al. <sup>[32]</sup>	Framework development study	Not applicable (simulation-based dataset, n=30)	IoT+ML (SVM-based classification)	Monitoring system	High activity classification accuracy (~98.8%) with real-time tracking of diet and physiological data	Demonstrates the potential of integrated AI-IoT systems for continuous monitoring and personalized feedback
Finkelstein et al. <sup>[33]</sup>	Secondary analysis of RCT (DIRECT-PLUS)	Overweight/obese adults (n=216)	Deep learning (CNN ensemble)+explainable AI	Biomarker identification	Brain-derived BMI predictions correlated with weight loss; distinct neuroanatomical patterns identified	Demonstrates the potential of AI to identify neurobiological signatures of obesity and monitor intervention effects
Cooper et al. <sup>[34]</sup>	Cross-sectional study (Human Connectome Project)	Healthy adults (n=1106)	Deep learning (3D-CNN)+explainable AI	Neurobiological modeling	High BMI prediction accuracy (R2≈0.44); deep learning captured complex brain-obesity associations	Demonstrates the potential of AI to uncover brain-based biomarkers and advance precision obesity research

AI: Artificial intelligence; ML: Machine learning; DL: Deep learning; RL: Reinforcement learning; IoT: Internet of things; CGM: Continuous glucose monitoring; SVM: Support vector machine; CNN: Convolutional neural network; 3D-CNN: Three-dimensional convolutional neural network; k-MRI: Magnetic resonance imaging; BMI: Body mass index.

The proposed system begins with personal information, health history, dietary preferences, and lifestyle data obtained from the user. Food recognition and diet tracking are performed using image-based algorithms (e.g., convolutional neural network [CNN], YOLO, Faster R-CNN) and transfer learning models; ML methods analyze food diaries, questionnaires, and wearable device data. This assessment phase is followed by disease risk prediction and personalized nutrition recommendations; algorithms such as random forest, XGBoost, SVM, ensemble methods, NLP, and RL are used here. In addition, genetic analysis and optimization algorithms support personalized meal planning. The process continues through a feedback loop via mobile applications and wearable technologies, ensuring the system is dynamically updated. Ultimately, the proposed framework aims to deliver precise, adaptable, and sustainable personalized nutrition management by combining behavioral insights and multidimensional data analysis.<sup>[38]</sup>

Collectively, these findings suggest that AI facilitates a multi-dimensional understanding of obesity by integrating genetic, clinical, behavioral, and environmental data, thereby supporting precision-based prevention and personalized management strategies.<sup>[39,40]</sup>

### Limitations

One limitation of this study is the limited number of intervention studies, as well as the predominance of short-term and pilot designs in the existing literature, with a notable lack of long-term randomized controlled trials. This limits the ability to draw robust conclusions regarding the sustainability and long-term clinical effectiveness of AI-driven weight management interventions. Furthermore, many studies are still in the developmental phase. Therefore, no definitive conclusions can be drawn regarding the safety of participants or the effectiveness and accuracy of weight loss outcomes derived from artificial intelligence; these aspects require further evaluation. Another limitation is the language restriction, as studies not published in English were excluded.

In addition, one of the major shortcomings in nutrition research using artificial intelligence is the failure to properly address ethical issues, especially considering the potential for artificial intelligence to function as a dietitian. People should feel secure about how their data will be stored, analyzed, and used. Therefore, we recommend transparency when working with artificial intelligence in the field of nutrition

### Conclusion

AI-supported weight management approaches are rapidly evolving from exploratory digital tools to clinically meaningful support systems.

In the field of Nutrition and Dietetics, positioning AI at Level 3 (human-supported autonomy) offers a safer and more sustainable approach. This model supports the decision-making process by establishing a balanced structure between algorithmic efficiency and clinical responsibility while maintaining human oversight. In contrast, Level 4 fully autonomous clinical decision systems harbor significant uncertainties in terms of ethics, legal responsibility, and regulation. Particularly in the field of nutrition, where individual metabolic and behavioral differences are pronounced, fully autonomous systems carry the risk of leading to erroneous generalizations and undesirable clinical outcomes.

The most sustainable and ethically sound model indicated by the literature is a framework in which AI enhances clinical expertise rather than replacing dietitians. Decision support systems, real-time monitoring, and adaptive feedback mechanisms can increase personalization and scalability; however, algorithmic outputs must be evaluated within a clinical, cultural, and ethical context. Professional oversight is indispensable, especially considering risks such as algorithmic bias, data privacy, and misinterpretation of context-disconnected recommendations.

Future research should prioritize large-scale, multi-center randomized controlled trials, external validation of AI models across diverse populations, cost-effectiveness analyses, and the development of clear regulatory frameworks to ensure safe and ethical implementation.

**Ethics Committee Approval:** As this study is a narrative review based on previously published literature, ethics committee approval was not required.

**Conflict of Interest:** None declared.

**Financial Disclosure:** The author declared that this study received no financial support.

**Use of AI for Writing Assistance:** The author used artificial intelligence (AI)-assisted tools (e.g., ChatGPT Plus) to improve the clarity and readability of the manuscript. The content was critically reviewed, revised, and validated by the author. The author takes full responsibility for the accuracy, integrity, and originality of the work.

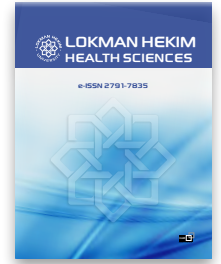
**Acknowledgments:** The author would like to express sincere gratitude to Prof. Dr. Ayla Gülden Pekcan for shaping the selection of this topic.

**Peer-review:** Double blind peer-reviewed.

## References

1. Tsolakidis D, Gymnopoulos LP, Dimitropoulos K. Artificial intelligence and machine learning technologies for personalized nutrition a review. *Informatics* 2024;11(3):62. [CrossRef]
2. Cruz Fernandez M, Castillo-Velásquez FA, Rodriguez-Abreo O, Ortiz-Moctezuma E, Iturralde Carrera LA, Estévez-Bén AA, et al. Neural network-based weight loss prediction: behavioral integration of stress and sleep in AI decision support. *AI* 2025;6(9):210. [CrossRef]
3. WHO. WHO acceleration plan to stop obesity. In: World Health Organization. 2023. Available at: <https://iris.who.int/server/api/core/bitstreams/ea789198-6336-45e3-adfd-7abb4b147b69/content> Accessed 6 Feb, 2026
4. Nolting M, Walsh G. Integrating AI in digital weight loss coaching: enhancing automated health guidance for scalable learning. *International Journal of Intelligent Computing and Cybernetics* 2025;18:523-36. [CrossRef]
5. World Obesity Federation. World Obesity Atlas. In: London: World Obesity Federation. (2025). <https://data.worldobesity.org/publications/world-obesity-atlas-2025-v7.pdf> Accessed 6 Feb, 2026.
6. Şenol H, Çolak H. Forecasting childhood and adolescent obesity prevalence in Türkiye through 2030: A novel application of artificial neural networks for public health planning. *Eng Appl Artif Intell* 2025;159:111720. [CrossRef]
7. Fu L, Burns R, Xie Y, Shen J, Zhe S, Estabrooks P, et al. The development and Use of AI chatbots for health behavior change: scoping review. *J Med Internet Res* 2026;28:e79677. [CrossRef]
8. Erik A, Hamidy SM, Karamancıoğlu H, Küçük Kırtıklı BN. AI in nutrition: multi-criteria analysis of diet plans across diverse client profiles. *Nutr Res* 2026;146:53-67. [CrossRef]
9. Lee V. Introduction to the dietary management of obesity in adults. *Clin Med (Lond)* 2023;23(4):304-10. [CrossRef]
10. Buttriss J, Hickson M, Whelan K, Williams C. Navigating the complexity of applying nutrition evidence to individualised care: Summary of an Academy of Nutrition Sciences Position Paper. *Nutr Bull* 2024;49(1):1-5. [CrossRef]
11. Boaventura B, Stanford FC. Dietitians' role in obesity care: insights from the Academy of Nutrition and Dietetics. *Hepatobiliary Surg Nutr* 2025;14(1):159-62. [CrossRef]
12. Johnson VR, Washington TB, Chhabria S, Wang EHC, Czepiel K, Reyes KJC, et al. Food as medicine for obesity treatment and management. *Clin Ther* 2022;44:671-81. [CrossRef]
13. Rugji J, Erol Z, Hamadani A, Gjoni Gündemir M, Taşçı F, Musa L. Next-generation nutrition: Innovative and AI-tailored concentrated ingredients. *Trends Food Sci Technol* 2026;168:105515. [CrossRef]
14. Chen S, Dai Y, Ma X, Peng H, Wang D, Wang Y. Personalized optimal nutrition lifestyle for self obesity management using metaalgorithms. *Sci Rep* 2022;12(1):12387. [CrossRef]
15. Wu X, Oniani D, Shao Z, Arciero P, Sivarajkumar S, Hilsman J, et al. A Scoping review of artificial intelligence for precision nutrition. *Adv Nutr* 2025;16(4):100398. [CrossRef]
16. Haghhighathoseini A, Lin S yu, Song G, An R. A systematic review on applications of artificial intelligence for obesity prevention. *Obesity Reviews* 2026;27(5):e70062. [CrossRef]
17. Krauss P. What is artificial intelligence? In: *Artificial Intelligence and Brain Research*. Berlin, Heidelberg: Springer; 2024. p. 107-12. [CrossRef]
18. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon* 2024;10(4):e26297. [CrossRef]
19. Konstantinidis D, Blekos A, Kotaidou M, Papaioannou P, Gymnopoulos L, Dimitropoulos K. Using AI for personalized nutrition: customizing diets for optimal consumer well-being. In: Alexander P, editor. *Encyclopedia of Agriculture and Food Systems*. 3<sup>rd</sup> ed. Academic Press 2026;3:80-95. [CrossRef]
20. Lacruz-Pleguezuelos B, Bazán GX, Romero-Tapiador S, Freixer G, Tolosana R, Daza R, et al. AI4Food, a feasibility study for the implementation of automated devices in the nutritional advice and follow up within a weight loss intervention. *Clin Nutr* 2025;48:80-9. [CrossRef]
21. Forman EM, Berry MP, Butryn ML, Hagerman CJ, Huang Z, Juarascio AS, et al. Using artificial intelligence to optimize delivery of weight loss treatment: Protocol for an efficacy and cost-effectiveness trial. *Contemp Clin Trials* 2023;124:107029. [CrossRef]
22. Huntriss R, Salimgaraev R, Nikogosov D. Older adults see highest levels of engagement with ai coach in a weight loss application. *J Acad Nutr Diet* 2025;125(10):A59. [CrossRef]
23. Abdelaal M, le Roux CW, Docherty NG. Effectiveness of an artificial intelligence-assisted app for improving eating behaviors: mixed methods evaluation. *J Med Internet Res* 2017;5(7)26.
24. Bhagavathula AS. Artificial intelligence and natural language processing of patient narratives to evaluate semaglutide for weight loss. *Ann Epidemiol* 2025;111:9-13. [CrossRef]
25. Brown C, Iyer A, Kumbara A, Ebert M. Nutrition analysis and advanced artificial intelligence (AI) predicts weight loss for people with type 2 diabetes (T2D). *J Acad Nutr Diet* 2024;124:A28. [CrossRef]
26. Wong WK, Chimoriya R, Acosta P, Kunte PS, Hardikar HP, Kormas N, et al. AI-driven approaches to identifying baseline characteristics of weight loss in a multidisciplinary metabolic weight management program for people with Class 3 obesity. *Obes Res Clin Pract* 2024;18:S58. [CrossRef]
27. Pokushalov E, Garcia C, Ponomarenko A, Samoilova I, Smith J, Johnson M, et al. Optimizing weight loss with artificial intelligence: A randomized controlled trial of dietary-supplement prescriptions in adults with overweight and obesity. *Clin Nutr ESPEN* 2025;69:75-86. [CrossRef]
28. Kamar M, Ryan DH, Leonard S, Wyatt HR, Kenan Y, Asaraf LC, et al. The safety and efficacy of extended use of an oral shape-shifting superabsorbent hydrogel capsule for weight loss: The ELECT extension study. *Obesity Pillars* 2025;16:100216. [CrossRef]

29. An R, Shen J, Xiao Y. Applications of artificial intelligence to obesity research: scoping review of methodologies. *J Med Internet Res* 2022;24(12):e40589. [\[CrossRef\]](#)
30. Haghghathoseini A, Lin S yu, Song G, An R, Xue H, et al. A systematic review on applications of artificial intelligence for obesity prevention. *Obesity Reviews* 2025;27:e70062. [\[CrossRef\]](#)
31. Saikia AP, Kalita A. Artificial Intelligence and Multi-Omics Integration in Obesity: A Review of Computational Models for Predicting Metabolic Comorbidities. *Clinical & Translational Metabolism* 2025;23:7. [\[CrossRef\]](#)
32. Alsareii SA, Raza M, Alamri AM, AlAsmari MY, Irfan M, Raza H, et al. Artificial intelligence and internet of things enabled intelligent framework for active and healthy living. *Computers, Materials & Continua* 2023;75:3833-48. [\[CrossRef\]](#)
33. Finkelstein O, Levakov G, Kaplan A, Zelicha H, Meir AY, Rinott E, et al. Deep learning-based BMI inference from structural brain MRI reflects brain alterations following lifestyle intervention. *Human Brain Mapping* 2024;45(3):e26595. [\[CrossRef\]](#)
34. Cooper A, Elsayed M, Owens MM, MacKillop J. Deep learning using structural MRI dramatically improves cross-validated prediction accuracy of body mass index. *Brain Mechanisms* 2026;151:202524. [\[CrossRef\]](#)
35. Anwar A, Rana S, Pathak P. Artificial intelligence in the management of metabolic disorders: a comprehensive review. *J Endocrinol Invest* 2025;48:1525-38. [\[CrossRef\]](#)
36. Al-Taie A, Hafida M, Abdulsattar M, El Mahmoud R. Scoping insights into artificial intelligence-driven treatment of diabetes mellitus in clinical practice. *Egypt J Intern Med* 2026;38:12. [\[CrossRef\]](#)
37. Joshi S, Bisht B, Kumar V, Singh N, Pasha SBJ, Singh N, et al. Artificial intelligence assisted food science and nutrition perspective for smart nutrition research and healthcare. *Systems Microbiology and Biomanufacturing* 2023;4:86-101. [\[CrossRef\]](#)
38. Theodore Armand TP, Nfor KA, Kim JI, Kim HC. Applications of artificial intelligence, machine learning, and deep learning in nutrition: a systematic review. *Nutrients* 2024;16(7):1073. [\[CrossRef\]](#)
39. Azmi S, Kunnathodi F, Alotaibi HF, Alhazzani W, Mustafa M, Ahmad I, et al. Harnessing artificial intelligence in obesity research and management: a comprehensive review. *Diagnostics* 2025;15(3):396. [\[CrossRef\]](#)
40. Chew HSJ, Chew NW, Loong SSE, Lim SL, Tam WSW, Chin YH, et al. Effectiveness of an artificial intelligence-assisted app for improving eating behaviors: mixed methods evaluation. *J Med Internet Res* 2024;26:e46036. [\[CrossRef\]](#)



## REVIEW

# The Governance of Artificial Intelligence in Healthcare: Ethical Foundations, Legal Challenges, and Implementation Realities

 **Mahmut Yılmaz**

Department of Intensive Care, İzmir City Hospital, İzmir, Türkiye

## Abstract

Artificial intelligence (AI) is transforming healthcare across diagnostics, decision-making, and clinical workflows, yet its integration raises complex ethical, legal, and operational challenges. This narrative review synthesizes three traditionally fragmented domains: Ethical principles, legal accountability, and implementation realities. We draw on literature from major databases alongside regulatory frameworks, including the World Health Organization, the Organisation for Economic Co-operation and Development, the National Institute of Standards and Technology, the European Union (EU), the Food and Drug Administration (FDA), and the International Medical Device Regulators Forum, and examine Türkiye's policies (e.g., Personal Data Protection Law No. 6698) to provide a middle-income country perspective. This review makes three contributions. First, we reconceptualize core bioethical principles – autonomy, beneficence, non-maleficence, and justice – in AI-mediated settings, emphasizing transparency, human oversight, and equity-sensitive design. Second, we frame legal accountability as a distributed system involving developers, institutions, and clinicians. Third, we bridge theory and practice through real-world cases (sepsis prediction vs. proprietary algorithms) and propose an integrated lifecycle governance model. Comparative analysis of the EU AI Act, FDA's 2026 guidance, and Türkiye's regulatory landscape shows convergence toward risk-based governance, alongside persistent gaps, particularly in middle-income settings. Responsible AI governance requires not only regulatory compliance but also continuous evaluation, transparency, and human-centered oversight. Despite global convergence on high-level principles, significant gaps remain in translating these into enforceable mechanisms and clinical practice. Future research should prioritize empirically validated governance models that ensure AI augments – rather than undermines – clinical judgment and patient trust.

**Keywords:** Artificial intelligence; Clinical implementation; Governance; Healthcare ethics; Legal accountability; Regulatory frameworks; Türkiye

**A**rtificial intelligence (AI) has moved from an experimental technology to an increasingly routine component of healthcare systems. It now shapes diagnostic pathways, prognostic assessments, clinical decision-support tools, and even the administrative and logistical functions of health institutions. The appeal of AI – particularly machine learning

models and, more recently, large language models – lies in its ability to process complex data, recognize patterns at scale, and potentially tailor care to individual patients. However, healthcare is not a neutral technological environment. Decisions made in clinical settings carry immediate and sometimes irreversible consequences, meaning that errors,

**Cite this article as:** Yılmaz M. The Governance of Artificial Intelligence in Healthcare: Ethical Foundations, Legal Challenges, and Implementation Realities. *Lokman Hekim Health Sci* 2026;6(2):372–382.

**Correspondence:** Mahmut Yılmaz, M.D. İzmir Şehir Hastanesi, Yoğun Bakım Kliniği, İzmir, Türkiye

**E-mail:** mahmutyilmazmd@gmail.com **Submitted:** 01.05.2026 **Revised:** 01.05.2026 **Accepted:** 04.05.2026 **Available Online:** 16.06.2026



**OPEN ACCESS** This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).



hidden biases, or opaque decision-making processes can directly translate into patient harm. For this reason, AI in healthcare must be understood not simply as a technical advancement but as a socio-technical intervention that reshapes patient rights, professional responsibilities, and institutional accountability. International guidance increasingly stresses that the benefits of AI can only be sustained when systems are governed with explicit attention to safety, effectiveness, transparency, and responsible oversight throughout their lifecycle.<sup>[1,2]</sup>

Recent discussions on trustworthy AI in healthcare have further emphasized that adoption is increasingly recognized to be slowed not by technical limitations but by unresolved ethical concerns. Issues such as algorithmic bias, lack of transparency, privacy risks, and ambiguous accountability structures continue to undermine trust. These concerns point toward the need for multidimensional governance models that integrate fairness, robustness, safety, transparency, privacy, and accountability as interconnected rather than isolated principles.<sup>[3,4]</sup> Recent meta-analyses of global AI ethics guidelines further demonstrate a growing convergence around core principles – including transparency, justice, non-maleficence, responsibility, and privacy – while simultaneously highlighting substantial variation in their implementation across jurisdictions and sectors.<sup>[5]</sup> Similarly, unified ethical frameworks have proposed adding explicability as a core principle, linking transparency directly to accountability and interpretability in algorithmic systems.<sup>[6]</sup> At the same time, governance discussions remain heavily concentrated on European Union (EU) and United States (US) regulatory discourse, whereas implementation in middle-income jurisdictions often occurs within fragmented legal and institutional environments.

This review, therefore, synthesizes recent literature and regulatory documents to examine the ethical foundations, legal accountability structures, and emerging governance approaches for responsible AI integration into healthcare systems.

## Materials and Methods

This narrative review critically synthesized contemporary ethical and legal challenges associated with the implementation of AI across healthcare systems. The review was designed as a structured narrative review rather than a systematic review or meta-analysis. Accordingly, its objective was not the exhaustive retrieval of all published studies, but the identification and critical synthesis of conceptually influential, regulatorily authoritative, and clinically relevant literature.

We searched the following electronic databases: PubMed/MEDLINE, Scopus, and Google Scholar using a Boolean search strategy developed across four conceptual domains: AI technologies, healthcare context, ethical principles, and legal/regulatory governance. The structured database search focused on literature published between January 2021 and February 2026. To ensure conceptual completeness, seminal earlier publications identified through hand-searching and backward reference-list screening were also incorporated when they were considered foundational to the ethical, legal, and governance discourse on AI in healthcare.

### The Boolean Search Syntax Used was as Follows

("Artificial intelligence" OR "machine learning" OR "deep learning" OR "clinical decision support" OR "generative AI" OR "large language model") AND (healthcare OR medicine OR clinical) AND (ethics OR fairness OR bias OR accountability OR transparency OR explainability OR privacy OR justice) AND (law OR regulation OR governance OR liability OR "risk management" OR "medical device regulation" OR "malpractice").

Inclusion criteria comprised peer-reviewed original research, policy analyses, regulatory frameworks, consensus reports, and governance guidelines addressing the ethical or legal implications of AI in healthcare delivery. Articles limited exclusively to technical algorithm development without substantial ethical or regulatory discussion were excluded. To ensure the currency and authority of the synthesized material, particular emphasis was placed on literature published in high-impact, peer-reviewed journals and on documents issued by authoritative international bodies, including the World Health Organization (WHO), the Organization for Economic Co-operation and Development (OECD), the European Commission, and the National Institute of Standards and Technology (NIST). The reference lists of eligible publications were additionally screened to identify further relevant institutional documents. Consistent with narrative review methodology, conceptual relevance, regulatory authority, and contemporary applicability were prioritized over quantitative synthesis. No formal preferred reporting items for systematic reviews and meta-analyses flow diagram, duplicate screening workflow, or risk-of-bias assessment tool was applied, because the review was not designed as a systematic evidence synthesis. This methodological choice was deliberate: The aim was not to quantify effect sizes or aggregate study outcomes, but to critically synthesize

conceptually influential and regulatorily authoritative sources across a rapidly evolving field where randomized controlled trials remain scarce.

Accordingly, this review synthesizes selected literature and regulatory materials to examine the core ethical principles, legal accountability structures, and emerging governance approaches guiding responsible AI integration into healthcare systems.

## Ethical and Regulatory Analysis of AI in Healthcare

The foundational bioethical principles – autonomy, beneficence, non-maleficence, and justice – remain central to medical practice, yet their practical meaning shifts when clinical decisions are influenced by algorithmic systems. Autonomy, for instance, becomes difficult to safeguard when patients are asked to consent to recommendations generated by models whose internal logic may be inaccessible even to clinicians. In such contexts, informed consent risks becoming a procedural formality rather than a meaningful exchange of information. Respecting autonomy in algorithm-mediated care, therefore, requires more than technical accuracy; it demands transparency about the role of AI, honest communication about uncertainty, and the preservation of a human decision-maker capable of contextualizing algorithmic outputs within the patient's personal values and clinical circumstances. WHO guidance reinforces this perspective by emphasizing that humans must remain in control of medical decisions, rather than delegating authority to opaque systems.<sup>[2]</sup>

The principles of beneficence and non-maleficence similarly acquire new dimensions. AI systems are often justified on the grounds that they improve efficiency or diagnostic accuracy, yet these benefits must be weighed against new categories of harm. Dataset shift, limited external validity, spurious correlations, and automation bias can all lead to inappropriate clinical decisions. In particular, automation bias may encourage clinicians to defer to algorithmic outputs even when these conflict with clinical intuition. Another emerging concern is alert fatigue, where excessive algorithm-generated warnings may desensitize clinicians and increase the likelihood of critical oversights. Such risks become especially pronounced when models are deployed in populations or environments different from those used in training, or when performance deteriorates over time. A safety-oriented ethical approach, therefore, requires continuous evaluation, real-world validation, and clear mechanisms for retraining or withdrawal. WHO regulatory considerations highlight safety,

effectiveness, and stakeholder dialogue as key elements in balancing risks and benefits.<sup>[1]</sup>

Justice and fairness are perhaps the most complex ethical challenges in AI-mediated healthcare. Clinical data often reflect longstanding disparities in access to care, diagnosis, and treatment. When training datasets fail to adequately represent certain populations, models may systematically underperform for those groups, thereby reinforcing existing inequities. Contemporary analyses have already documented such patterns in clinical AI systems, with potential implications for unequal treatment outcomes.<sup>[7]</sup> This concern is strongly supported by empirical evidence demonstrating that widely used healthcare algorithms may systematically disadvantage certain populations; for example, a large-scale study revealed that a commonly used risk prediction algorithm significantly underestimated the health needs of Black patients due to biased proxy variables.<sup>[8]</sup> Such findings illustrate that algorithmic bias is not merely theoretical but a measurable, clinically relevant source of inequity. It is therefore important to distinguish between algorithmic bias, which originates from technical flaws in data or model design, and the broader concept of fairness, which concerns the just distribution of healthcare benefits and burdens. Bias may arise not only from incomplete datasets but also from proxy variables, measurement errors, or structural inequities embedded in healthcare systems. Ethical deployment of AI thus requires subgroup performance evaluation, transparent reporting of limitations, and proactive mitigation strategies. WHO guidance explicitly identifies inclusiveness and equity as core ethical imperatives.<sup>[2]</sup>

Transparency and explainability occupy a central position at the intersection of ethics, law, and clinical practice. Although black-box models may achieve impressive predictive accuracy, their lack of interpretability can undermine shared decision-making and weaken professional accountability. Clinicians cannot reasonably justify decisions they do not understand, nor can patients provide meaningful consent in the absence of intelligible explanations. From a legal perspective, this challenge aligns closely with emerging data protection principles, such as the "right to explanation," which restricts fully automated decision-making and requires individuals to obtain meaningful information about algorithmic outputs.<sup>[9]</sup> This reinforces the necessity of interpretability not only as an ethical requirement but also as a developing legal standard. Trustworthy AI frameworks, therefore, identify transparency, explainability, privacy, fairness, and accountability as essential dimensions for safe implementation.<sup>[3]</sup> This emphasis is echoed in governance initiatives such as the NIST AI risk management

framework, which characterizes trustworthy AI through attributes including validity, reliability, safety, transparency, explainability, accountability, and privacy.<sup>[10]</sup>

Data privacy and confidentiality have long been central ethical obligations in healthcare, but the data-intensive nature of AI introduces additional complexities. The development and deployment of AI systems typically require large volumes of clinical and imaging data, both of which are highly sensitive and subject to strict regulatory protections. Healthcare institutions must therefore navigate complex legal frameworks designed to safeguard personal health information and ensure responsible data use.

The increasing reliance on cloud-based infrastructures and cross-border data transfers further complicates this landscape by introducing jurisdictional challenges and potential regulatory conflicts.<sup>[2,3,11]</sup> Legal instruments such as the Health Insurance Portability and Accountability Act and the general data protection regulation (GDPR) establish safeguards for personal health information, mandate lawful data processing, and restrict certain forms of automated decision-making. Ethical oversight mechanisms, including Institutional Review Boards, also play a crucial role in assessing risks, ensuring informed consent, and protecting privacy and fairness in AI-related healthcare research.<sup>[11]</sup> Ethical data governance, therefore, requires data minimization, clear purpose limitation, secure storage, and transparency regarding secondary data use. WHO guidance warns that inadequate safeguards may erode public trust and ultimately result in patient harm.<sup>[1,2]</sup>

## Legal Accountability and Regulatory Frameworks: International and National Perspectives

Questions of legal accountability represent another major challenge. AI systems diffuse causal responsibility across multiple actors, including developers, vendors, and healthcare institutions, clinicians, and data providers. When harm occurs, identifying the responsible party becomes significantly more complex than in traditional malpractice scenarios. Conventional liability models focus primarily on clinician negligence, yet AI-related harm may stem from flawed model design, insufficient validation, poor integration into clinical workflows, or inadequate performance monitoring. A pragmatic approach is therefore to conceptualize accountability as shared but clearly delineated, with each stakeholder assuming responsibilities appropriate to their role. This perspective aligns with the OECD principles on AI accountability.<sup>[12,13]</sup>

## International Regulatory Frameworks

Regulatory frameworks are evolving in response to these challenges. In the EU, the Artificial Intelligence Act (AI Act) entered into force in August 2024, introducing a risk-based approach to AI regulation. Its application is phased, with some provisions taking effect from February 2025 and the main body of obligations from August 2026. Many medical AI applications are classified as high-risk systems, subject to stringent requirements concerning training data quality, conformity assessments, risk management, transparency obligations, and human oversight.<sup>[14,15]</sup>

Additional guidance clarifies how the AI Act interacts with existing medical device regulations, underscoring the need for coordinated compliance across regulatory regimes.<sup>[16]</sup> The interplay between the EU AI Act and established medical device regulatory frameworks further complicates compliance processes by introducing overlapping obligations related to risk classification, data governance, and post-market monitoring, highlighting emerging challenges of regulatory fragmentation in AI-enabled healthcare systems.<sup>[17]</sup> The phased implementation timeline further highlights the importance of aligning institutional deployment strategies with regulatory milestones.

In the US, regulatory oversight largely depends on whether a software function is classified as a medical device. The Food and Drug Administration (FDA) guidance on clinical decision-support software distinguishes between exempt and regulated systems based on functionality and transparency criteria.<sup>[18]</sup> Meanwhile, the concept of Software as a Medical Device, developed by the International Medical Device Regulators Forum, provides a widely accepted framework for risk categorization and regulatory approaches.<sup>[19]</sup> Ethical oversight bodies, such as Institutional Review Boards, continue to play a central role in ensuring that AI research involving human data complies with standards for informed consent, privacy protection, and risk minimization.<sup>[11]</sup>

## National Regulatory Perspective: The Case of Türkiye

While the majority of international governance discourse focuses on the EU and the US, a comprehensive legal analysis requires examination of national frameworks in other jurisdictions. Türkiye represents an instructive case, as it has developed a multi-layered regulatory approach to AI in healthcare without yet adopting a comprehensive AI-specific law comparable to the EU AI Act.<sup>[20-23]</sup>

First, data protection constitutes the foundational layer. Türkiye's Personal Data Protection Law No. 6698, enacted

in 2016 and substantially amended in 2024 to align more closely with GDPR, governs the processing of personal data, including health data.<sup>[21,23,24]</sup> Under Article 6 of the Personal Data Protection Law No. 6698, health data are classified as special categories of personal data and are subject to heightened protection.<sup>[24]</sup> For AI development, this means that training datasets derived from electronic health records or other clinical repositories require either patient consent or a lawful processing basis under the law. Importantly, Article 11 of the same law grants data subjects the right to object to a result arising against them through analysis performed solely by automated systems; however, the law does not establish a fully articulated AI-specific “right to explanation” comparable to broader debates under the GDPR framework.<sup>[21,23,24]</sup>

Second, medical device regulation provides the product safety layer. The Turkish Medicines and Medical Devices Agency has adopted the Medical Device Regulation (MDR) (harmonized with EU 2017/745), which defines software used for medical purposes as a medical device.<sup>[21–23,25]</sup> Consequently, AI-based software intended for diagnostic or therapeutic purposes is classified as an active medical device.<sup>[21–23,25]</sup> The risk classification rules (Annex VIII of the MDR) categorize most clinical AI applications as Class IIa or IIb devices, which require conformity assessment by notified bodies.<sup>[21,23,25]</sup> However, unlike the EU AI Act, Türkiye’s MDR does not yet impose specific requirements for training data quality, bias mitigation, or human oversight beyond general safety and performance requirements.<sup>[21–23]</sup>

Third, digital health infrastructure is governed by the Regulation on the Provision of Remote Healthcare Services.<sup>[26]</sup> This regulation primarily structures the legal framework for remote healthcare delivery and related digital service processes. Although it is relevant to the broader digital health ecosystem, it does not expressly provide a dedicated framework for AI model development, algorithm validation, or post-deployment monitoring. This creates uncertainty for developers and institutions seeking to access and use real-world clinical data for AI innovation within a clearly defined regulatory pathway.<sup>[21,23,26]</sup>

Recent Turkish scholarship has also emphasized that AI applications in healthcare – particularly in high-risk contexts such as pre-hospital emergency services – raise unresolved legal questions regarding malpractice liability, patient autonomy, and the protection of sensitive health data.<sup>[27]</sup> These analyses underscore that, despite existing regulatory instruments, there remains a need for more detailed, AI-specific legal guidance addressing autonomous or semi-autonomous clinical decision-making systems.

Fourth, liability and malpractice law in Türkiye has not yet specifically addressed AI-related harm. Under the Turkish Code of Obligations (Law No. 6098), medical malpractice liability is primarily fault-based.<sup>[20,21,23,28]</sup> In cases involving AI, potential liable parties include: The clinician (for over-reliance on erroneous AI output), the hospital (for inadequate validation or training), and the AI developer (for product defect).<sup>[20,21,23,28]</sup> However, no binding Turkish court precedent has yet established apportionment of liability in AI-related medical injury.<sup>[20,21,23]</sup> Legal scholars have noted this gap, calling for either legislative clarification or judicial development of AI-specific liability rules.<sup>[20,21,23]</sup>

Table 1 provides a comparative overview of regulatory approaches to AI in healthcare across the EU, the US, and Türkiye.

## Human Oversight as a Cross-Cutting Principle

Across ethical guidelines and regulatory frameworks, human oversight remains a consistent and central principle. AI systems are intended to assist, rather than replace, clinical decision-making. However, meaningful oversight requires more than nominal human involvement. Clinicians must be adequately trained, workflows must be transparent, and systems must include mechanisms for auditing and review. Trustworthy AI frameworks emphasize human-in-the-loop approaches, auditability, and traceability as essential governance tools.<sup>[3]</sup> European regulatory analyses also suggest that healthcare institutions and clinicians deploying high-risk AI systems may bear specific compliance obligations, including monitoring and documentation responsibilities.<sup>[14,16,22]</sup>

## Implementation Realities: Case-Based Analysis of Clinical Deployment

While normative, ethical, and legal frameworks are essential, the title’s promise of “implementation realities” requires examination of actual clinical deployments. This section analyzes two real-world cases – one successful and one problematic – to derive concrete governance lessons.

### Case 1: Successful Deployment – Sepsis Prediction at Duke University Health System

Sendak et al.<sup>[29]</sup> (2020) documented the integration of a deep learning-based sepsis prediction model (the “Sepsis Watch”) into routine clinical care at Duke University Health System. The model was designed to identify patients at risk of sepsis up to 12 h before clinical recognition. Key governance elements that contributed to success included:

**Table 1.** Comparative regulatory approaches to AI in healthcare

Dimension	European Union (AI Act)	United States (FDA)	Türkiye (Personal Data Protection Law No. 6698 + Turkish Medicines and Medical Devices Agency)
Risk classification	High-risk where AI is itself a regulated medical device or a safety component of one, generally requiring third-party conformity assessment	Risk-based; classified as medical device if intended for diagnosis/treatment; pathways: 510(k), De Novo, or Premarket Approval	Class IIa/IIb under MDR (harmonized with EU); no AI-specific statutory risk tier yet
Transparency obligations	Article 13–technical documentation and user information	CDS transparency is relevant to whether the software remains outside the device definition or becomes subject to FDA oversight	Article 11 of the Personal Data Protection Law No. 6698–right to object to automated processing outcomes; no explicit “right to explanation.”
Human oversight	Mandatory for high-risk systems (Article 14)	Context-dependent and function-specific; not framed as a horizontal AI-law obligation comparable to the EU AI Act	Not explicitly required in MDR or Personal Data Protection Law No. 6698
Data quality requirements	Training/validation data must be relevant, representative, and bias-free	General software validation and quality system expectations	General data protection and medical device safety obligations; no AI-specific statutory dataset quality requirements
Liability/malpractice	No standalone AI-specific civil liability regime currently in force; existing product liability and national tort frameworks remain relevant. The proposed revision of the EU Product Liability Directive (2024) explicitly includes software and AI systems, which may reshape liability allocation for medical AI.	Traditional malpractice+product liability	Turkish Code of Obligations No. 6098 (fault-based); no AI-specific judicial precedent; scholarly debate ongoing. <sup>[15,16,18]</sup>
Post-market monitoring	Mandatory for high-risk systems	Post-market obligations arise through medical device oversight, where applicable	General vigilance under MDR; not AI-specific

SaMD: Software as a medical device; CDS: Clinical decision support; MDR: Medical device regulation; AI: Artificial intelligence; MDR: Medical device regulation; EU: European Union; FDA: Food and Drug Administration.

- Prospective validation: The model was validated on local data before deployment, addressing dataset shift concerns.
- Clinical workflow integration: A dedicated rapid response team received alerts, rather than adding notifications to already-overburdened physicians.
- Human-in-the-loop design: Nurses reviewed each alert before team activation, maintaining human judgment as a filter.
- Continuous monitoring: Performance was tracked weekly, with predefined thresholds for retraining or suspension.
- Clinician training: All users completed simulation-based training on AI limitations and appropriate override conditions.

From a legal accountability perspective, Duke established

a clear governance structure: The health system assumed responsibility for model validation and monitoring, while clinicians remained accountable for final treatment decisions. This delineation reduced ambiguity about liability in the event of adverse outcomes.

**Case 2: Problematic Deployment – Proprietary Sepsis Algorithm (Epic)**

In contrast, Wong et al.<sup>[30]</sup> (2021) published an external validation study of Epic Systems’ proprietary sepsis prediction model (the Epic Sepsis Model, ESM), which was deployed at hundreds of US hospitals. The study found that the model performed poorly on external validation, with an area under the receiver operating characteristic curve of 0.60–0.64, only marginally better than chance. More concerning, the model generated alerts for only 7% of sepsis patients while producing high false-positive rates.

**Table 2.** Comparative governance lessons from real-world AI deployment cases

Governance function	Successful (Duke Sepsis Watch)	Problematic (Epic ESM)
Local validation	Mandatory before deployment	Often skipped or superficial
Transparency	Model details shared with users	Proprietary, black-box
Human oversight	Nurse filter + team activation	Direct alert to physician
Performance monitoring	Weekly, with suspension thresholds	None documented
Liability clarity	Defined (hospital + clinician)	Ambiguous (vendor vs. hospital vs. clinician)

This table is derived from the case analyses presented in the Implementation Realities section.

### Ethical Failures Identified Included

- Lack of transparency: Epic did not disclose detailed model architecture or validation data, preventing independent assessment.
- Potential reinforcement of automation bias: Clinicians at deploying hospitals reported feeling pressured to respond to alerts despite low positive predictive value.
- Inadequate local validation: Many hospitals deployed the model without rigorous local performance testing.
- Absence of post-market surveillance: No systematic tracking of missed sepsis cases or false alarm harms.

Legal implications from this case include potential product liability claims against Epic (for defective software) and institutional negligence claims against hospitals (for failing to validate before deployment). The case illustrates how proprietary black-box models can undermine the ethical principle of transparency and create diffuse accountability. Table 2 provides a comparative summary of governance lessons derived from the Duke Sepsis Watch.<sup>[29]</sup> and the Epic Sepsis Model<sup>[30]</sup> case analyses. Together, these cases demonstrate that responsible AI governance is not merely about compliance with regulatory checklists, but about embedding continuous evaluation, transparency, and human judgment into clinical workflows.

### Special Populations and Clinical Domains

Certain patient groups and clinical domains raise additional ethical concerns. Pediatric populations, for example, rely on proxy consent and are frequently under-represented in training datasets, increasing the likelihood of diagnostic inaccuracies.<sup>[2,4,31]</sup>

Rare diseases and genomic medicine pose additional challenges, including the difficulty of scaling phenotype data collection and the presence of complex phenotypic overlap across distinct syndromes, which may complicate accurate diagnosis.<sup>[32]</sup> Emerging applications of AI in genomic diagnostics highlight both the potential to improve diagnostic performance and the importance of

carefully addressing limitations in data quality, bias, and generalizability.<sup>[32]</sup>

In mental health contexts, the use of AI for suicide risk prediction or sentiment analysis introduces complex questions about autonomy, stigma, and the therapeutic relationship.<sup>[2,4]</sup> Recent literature also emphasizes that algorithmic predictions in psychiatry may lack contextual interpretability and risk reinforcing existing biases, particularly when trained on non-representative behavioral or social data.<sup>[33]</sup> In such settings, ethical governance requires cautious deployment, a clearly defined scope, and enhanced monitoring mechanisms.

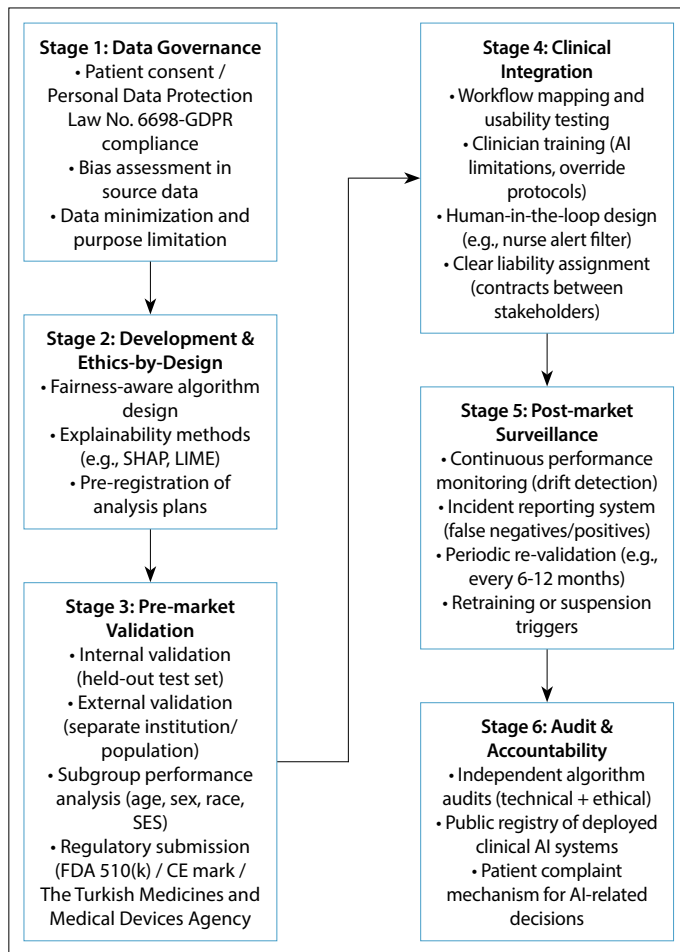
### A Lifecycle Governance Model for Responsible AI Implementation

Responsible AI implementation should be understood as a continuous lifecycle rather than a one-time technical intervention. This lifecycle includes ethics-by-design, rigorous validation, ongoing monitoring, and strong institutional governance. Trustworthy AI frameworks propose measurable dimensions – such as fairness, robustness, privacy, explainability, and accountability – to guide this process.<sup>[3]</sup>

In parallel, governance-oriented studies have proposed structured frameworks that operationalize these principles through lifecycle-based oversight, emphasizing bias mitigation, data governance, privacy protection, and accountability as core implementation domains.<sup>[34]</sup> Complementary analyses further highlight that trustworthiness in medical AI is shaped by interconnected factors, including data quality, algorithmic opacity, safety, and responsibility attribution, all of which require coordinated ethical and regulatory oversight.<sup>[35]</sup>

Figure 1 presents a synthesized governance model derived from the ethical principles, regulatory requirements, and implementation case analyses discussed above.

Practical implementation requires risk management strategies, thorough documentation, subgroup



**Figure 1.** Lifecycle governance model for ai in healthcare. This figure presents a synthesized lifecycle-based governance model for artificial intelligence in healthcare. Derived from the ethical principles, regulatory frameworks, and real-world implementation challenges discussed in this review, the model conceptualizes AI oversight across six sequential yet interdependent stages: Data Governance, Development and Ethics-by-Design, Pre-market Validation, Clinical Integration, Post-market Surveillance, and Audit and Accountability. The framework emphasizes that safe and equitable clinical AI requires continuous feedback between technical validation, regulatory compliance, human oversight, and post-deployment accountability.

AI: Artificial intelligence; FDA: U.S. Food and Drug Administration; GDPR: General data protection regulation; LIME: Local interpretable model-agnostic explanations; SES: Socioeconomic status, SHAP: Shapley additive explanations; WHO: World Health Organization.

performance audits, user training, and structured incident-reporting mechanisms. Frameworks such as the NIST AI risk management framework provide practical guidance for institutional governance, while binding regulations such as the EU AI Act establish minimum compliance obligations.<sup>[10,14,16]</sup> Ultimately, AI should serve as a tool that enhances, rather than replaces, clinical judgment, while preserving patient autonomy, equity, and accountability in healthcare delivery.<sup>[1,2]</sup>

## Clinical Implications for Healthcare Professionals and Institutions

Beyond theoretical ethical principles, AI-related risks materialize within everyday clinical workflows and institutional decision structures. Healthcare professionals increasingly interact with probabilistic algorithmic outputs that may influence diagnosis, treatment prioritization, or risk prediction. Ethical implementation, therefore, requires explicit documentation of whether AI-generated recommendations contributed to clinical decisions and how disagreements between clinician judgment and algorithmic output were resolved.

From a legal perspective, institutions deploying AI systems should establish clearly delineated accountability frameworks involving developers, vendors, healthcare organizations, and clinical end-users.<sup>[13]</sup> Based on the case analyses presented above, specific actionable recommendations include:

- For clinicians: Document when AI recommendations are overridden or followed; complete mandatory training on AI limitations and automation bias; report all AI-related adverse events or near-misses.<sup>[2,16]</sup>
- For institutions: Establish an AI governance committee with clinical, legal, and technical membership; mandate local validation before any clinical deployment; implement continuous performance dashboards with predefined suspension thresholds; secure liability agreements with AI vendors.<sup>[16,29]</sup>
- For developers: Provide model cards<sup>[36]</sup> detailing training data, performance subgroups, and intended use limitations; enable audit logging of all predictions; support explainability methods accessible to clinicians.<sup>[16]</sup>

In the Turkish context, institutions should additionally ensure compliance with Article 6 of the Personal Data Protection Law No. 6698 for health data processing, register AI-based medical devices with the Turkish Medicines and Medical Devices Agency, and consider obtaining legal opinions on liability allocation given the absence of AI-specific jurisprudence.<sup>[20,21]</sup>

Human oversight should not be interpreted as passive supervision but rather as an operational governance function supported by clinician training, institutional AI committees, and continuous performance evaluation.<sup>[2]</sup> Such measures reduce automation bias, maintain professional accountability, and align healthcare deployment practices with emerging international regulatory expectations.<sup>[14,16,17]</sup>

## Limitations

This narrative review has several limitations that should be acknowledged. First, the selection of sources involves inherent subjectivity; despite a structured search strategy with predefined inclusion criteria, relevant studies – particularly those reporting null or negative findings – may have been omitted. Second, the analysis predominantly reflects regulatory frameworks and implementation cases from the EU and the US, with only initial exploration of the Turkish context and limited coverage of other regions (e.g., Asia, Latin America, Africa). Third, the inclusion of non-peer-reviewed institutional documents (e.g., WHO guidance, NIST frameworks, national regulations) – while essential for capturing current regulatory thinking and authoritative governance standards – means that some sources have not undergone traditional peer review. Where possible, priority was given to documents with transparent development processes and public consultation mechanisms. Fourth, the rapidly changing legal landscape (e.g., phased implementation of the EU AI Act, ongoing development of Turkish digital health governance, and amendments to the Personal Data Protection Law No. 6698) means that some regulatory details may require updating. Fifth, the narrative review format does not permit quantitative synthesis or meta-analysis of implementation outcomes. Sixth, because the review was not conducted as a systematic review, no formal study-quality appraisal or evidence-certainty assessment was undertaken. Finally, the exclusion of non-English publications (except for Turkish regulatory documents) may have introduced language bias. Despite these limitations, the synthesis provides a structured foundation for understanding the governance challenges of AI in healthcare and identifying priority areas for future empirical research.

## Conclusion

AI is rapidly transforming healthcare by improving diagnostic accuracy, operational efficiency, and the personalization of care. Yet its integration into clinical practice also introduces complex ethical and legal challenges related to autonomy, fairness, transparency, privacy, and accountability. Addressing these challenges requires not only robust ethical frameworks but also clearly defined liability structures and comprehensive regulatory oversight.

International initiatives – including WHO guidance, the EU AI Act, and FDA regulatory approaches – reflect a growing global consensus that AI in healthcare must be governed responsibly. As this review demonstrates through a comparative analysis of EU, US, and Turkish frameworks, national regulatory contexts differ substantially, creating

challenges for multinational deployment and underscoring the need for context-sensitive governance. The analysis of real-world implementation cases (Duke Sepsis Watch vs. Epic Sepsis Model) further reveals that responsible governance requires not only regulatory compliance but also local validation, transparency, human oversight, and continuous monitoring. Specifically, Turkish policymakers and healthcare institutions should prioritize: (a) Developing binding guidance on how Article 11 of the Personal Data Protection Law No. 6698 (“right to object to automated outcomes”) applies in clinical settings; (b) clarifying the Turkish Medicines and Medical Devices Agency’s oversight role for AI as a medical device beyond initial conformity assessment; and (c) establishing a national AI incident reporting system for healthcare to generate the empirical evidence needed for future liability frameworks. The effectiveness of these efforts will depend not solely on regulatory compliance but also on fostering a culture of ethical awareness and interdisciplinary collaboration among clinicians, regulators, developers, and patients. Ultimately, AI should function as a tool that augments, rather than replaces, human clinical judgment while preserving the fundamental ethical principles of medical practice.

Future research should therefore move beyond normative ethical discussions toward empirically grounded governance models capable of evaluating real-world AI performance, safety, and accountability within clinical environments. Priority areas include: (1) Prospective studies of AI implementation measuring both clinical outcomes and unintended harms; (2) legal analyses of liability allocation in actual malpractice cases involving AI; (3) comparative effectiveness research on different governance models (e.g., centralized vs. distributed oversight); and (4) development of validated audit tools for algorithmic fairness in healthcare contexts, including in middle-income countries like Türkiye.

**Ethics Committee Approval:** Ethical approval was not required for this study since this is a review article.

**Conflict of Interest:** None declared.

**Financial Disclosure:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Use of AI for Writing Assistance:** During the preparation of this work, the author used artificial intelligence for language editing and grammar checking. After using this, the author reviewed and edited the content as needed and takes full responsibility for the publication’s content.

**Peer-review:** Double blind peer-reviewed.

## References

1. World Health Organization. WHO outlines considerations for regulation of artificial intelligence for health. Available at: <https://iris.who.int/server/api/core/bitstreams/ad62580f-540f-4e36-b957-e7f2946ae1fb/content> Accessed 12 Feb, 2026.
2. World Health Organization. Ethics and governance of artificial intelligence for health. Geneva: World Health Organization; 2021. Available at: <https://www.who.int/publications/item/9789240029200> Accessed 12 Feb, 2026.
3. Ahadian P, Xu W, Liu D, Guan Q. Ethics of trustworthy AI in healthcare: challenges, principles, and practical pathways. *Neurocomputing* 2026;661:131942. [CrossRef]
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44-56. [CrossRef]
5. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 2019;1(1):389-99. [CrossRef]
6. Floridi L, Cowls J. A unified framework of five principles for AI in society. *Harv Data Sci Rev* 2019;1(1):1-14. [CrossRef]
7. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022;28(1):31-8. [CrossRef]
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366(6464):447-53. [CrossRef]
9. Goodman B, Flaxman S. European Union regulations on algorithmic decision making and a "right to explanation". *AI Mag* 2017;38(3):50-7. [CrossRef]
10. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg (MD): U.S. Department of Commerce; 2023. Available at: <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf> Accessed 12 Feb, 2026.
11. Pantanowitz L, Hanna M, Pantanowitz J, Lennerz J, Henricks WH, Shen P, et al. Regulatory aspects of artificial intelligence and machine learning. *Mod Pathol* 2024;37(12):100609. [CrossRef]
12. Organisation for Economic Co-operation and Development. Recommendation of the Council on Artificial Intelligence. 2019. Available at: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> Accessed 12 Feb, 2026.
13. Cath C. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philos Trans A Math Phys Eng Sci* 2018;376(2133):20180080. [CrossRef]
14. European Commission. AI Act enters into force. Available at: [https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01\\_en](https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en) Accessed 12 Feb, 2026.
15. European Commission, Directorate-General for Health and Food Safety. Artificial intelligence in healthcare (EU eHealth and digital health & care). Available at: [https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare\\_en](https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en) Accessed 12 Feb, 2026.
16. Medical Device Coordination Group. MDCG 2025-6: Interplay between the Medical Devices Regulations (MDR/IVDR) and the Artificial Intelligence Act (AIA) 2025. Available at: [https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4\\_en](https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4_en) Accessed 12 Feb, 2026.
17. Kalodanis K, Feretzakis G, Rizomiliotis P, Verykios VS, Papapavlou C, Skrekas A, et al. Evaluating the impact of the EU AI act on medical device regulation. *Stud Health Technol Inform* 2025;323:40-4. [CrossRef]
18. U.S. Food and Drug Administration. Clinical decision support software: guidance for industry and Food and Drug Administration staff. Available at: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software> Accessed 12 Feb, 2026.
19. International Medical Device Regulators Forum. Software as a medical device (SaMD): key definitions. Available at: <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf> Accessed 12 Feb, 2026.
20. Dirî F. Yapay zeka teknolojisi ve beraberinde getirdiklerinin Türk sağlık hukuku kapsamında değerlendirilmesi. *Bilişim Hukuku Dergisi* 2024;6(1):270-320. [Article in Turkish] [CrossRef]
21. Yüzbaşıoğlu C. The use of artificial intelligence by the administration in health services and the resulting liability in Turkish law. *J Acad Res Med* 2025;15(3):109-11. [CrossRef]
22. Veale M, Zuiderveen Borgesius F. Demystifying the Draft EU Artificial Intelligence Act - analysing the good, the bad, and the unclear elements of the proposed approach. *Comput Law Rev Int* 2021;22(4):97-112. [CrossRef]
23. Günday HM, Albayrak Günday E. Artificial intelligence in mental health practices: legal liability analysis under Turkish, European, and common law frameworks. *Psikiyatride Güncel Yaklaşımlar* 2025;17:806-21. [CrossRef]
24. Republic of Türkiye. Personal Data Protection Law No. 6698. Official Gazette No. 29677. Available at: <https://mevzuat.gov.tr/mevzuat?MevzuatNo=6698&MevzuatTur=1&MevzuatTertip=5> Accessed 12 Feb, 2026.
25. Republic of Türkiye. Medical Device Regulation. Official Gazette No. 31499 (Reprint). Available at: <https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=38657&MevzuatTur=7&MevzuatTertip=5> Accessed 12 Feb, 2026. [in Turkish]
26. Ministry of Health of the Republic of Türkiye. Regulation on the provision of remote healthcare services. Official Gazette No. 31746, 10 February 2022. [cited 2026 Apr 02]. Available at: <https://www.resmigazete.gov.tr/eskiler/2022/02/20220210-2.htm> [in Turkish]
27. Şişli Z. Yapay zeka alanında hukuki gelişmeler ve hastane öncesi sağlık hizmetleri. *Medical Technologies Congress* 2024:117-20. [Article in Turkish]
28. Republic of Türkiye. Turkish Code of Obligations No. 6098. Official Gazette No. 27836. Available at: <https://resmigazete.gov.tr/eskiler/2011/02/20110204-1.htm> [Article in Turkish]
29. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform* 2020;8(7):e15182. [CrossRef]

30. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021;181(8):1065-70. [\[CrossRef\]](#)
31. Hua SBZ, Heller N, He P, Towbin AJ, Chen IY, Lu AX, et al. Underrepresentation of children in public medical imaging datasets. *Nature Health* 2026. [\[CrossRef\]](#)
32. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med* 2019;11(1):70. [\[CrossRef\]](#)
33. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, Jeste DV. Artificial intelligence for mental health and mental illnesses: an overview. *Curr Psychiatry Rep* 2019;21(11):116. [\[CrossRef\]](#)
34. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc* 2020;27(3):491-7. [\[CrossRef\]](#)
35. Zhang J, Zhang ZM. Ethics and governance of trustworthy medical artificial intelligence. *BMC Med Inform Decis Mak* 2023;23(1):7. [\[CrossRef\]](#)
36. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019; Atlanta, GA, USA. New York: ACM; 2019. p. 220-9. [\[CrossRef\]](#)



