

Review

Reporting Gaps and Challenges in Randomized Trials of Generative AI for Medical Use: Insights From a Systematic Review

Jae Eun Ko^{1*}, BEng; Ji Woo Hwang^{1*}, MEng; Sung Min Kim^{1,2,3}, PhD; Ji Yean Kwon^{1,2}, PhD

¹Department of Regulatory Science for Medical Device, Dongguk University, Seoul, Republic of Korea

²Department of Medical Device and Healthcare, Dongguk University, Seoul, Republic of Korea

³Department of Medical AI Engineering, College of Advanced Convergence Engineering, Dongguk University, Seoul, Republic of Korea

*these authors contributed equally

Corresponding Author:

Ji Yean Kwon, PhD

Department of Regulatory Science for Medical Device

Dongguk University

Pildong-ro 1gil 30, Jung-gu

Seoul 04620

Republic of Korea

Phone: 82 319615616

Email: jykwon@dgu.ac.kr

Abstract

Background: Generative AI (GenAI) is being deployed rapidly in medicine, and evidence of its clinical safety and efficacy depends on transparent reporting of randomized controlled trials (RCTs). The CONSORT-AI (Consolidated Standards of Reporting Trials–AI) extension is a benchmark for AI trial reporting, but it was drafted for conventional AI, and its applicability to GenAI has not been examined.

Objective: This study aimed to assess item-level adherence to the CONSORT-AI among RCTs evaluating GenAI for medical purposes and to distinguish deficiencies attributable to nonadherence from those attributable to items that do not correspond to how GenAI interventions are specified or delivered.

Methods: PubMed, CENTRAL (Cochrane Central Register of Controlled Trials), and Embase were searched for peer-reviewed, English-language RCTs published from 2022 to April 3, 2026, supplemented by backward and forward citation searching, in which patients, health care professionals, or laypeople received a GenAI intervention for a medical purpose; non-GenAI, nonmedical, nonrandomized, and non-peer-reviewed reports were excluded. Two reviewers independently screened studies, extracted data, and rated CONSORT-AI items as reported, partially reported, not reported, or not applicable using signaling questions, with disagreements resolved by a third reviewer (Cohen $\kappa=0.81$ and prevalence-adjusted, bias-adjusted $\kappa=0.84$). Risk of bias was not assessed, as the target was reporting adherence. Adherence was the proportion of reports rated “reported” per item; item-level adherence was compared across 9 characteristics using Fisher exact and Freeman-Halton tests (2-sided $\alpha=.05$).

Results: From 16,485 retrieved records, 74 reports from 73 RCTs (24,150 participants; median 100, IQR 50.0–247.0) were included. Large language models (LLMs) accounted for 98.6% (73/74) of interventions, with 60.8% (45/74) using GPT-family models. Item-level reporting adherence ranged from 0% (0/72) to 98.6% (73/74). High adherence occurred in general metadata and output specification items (range 59–73, 79.7%–98.6%). Conversely, reporting was poor for input and technical execution domains, input data eligibility (3/73, 4.1%), input acquisition/selection (9/74, 12.2%), algorithm versioning (17/74, 23.0%), trial setting integration (21/74, 28.4%), human-AI interaction (25/74, 33.8%), performance error (27/74, 36.5%), and open access/code (31/74, 41.9%). Input quality handling (Item 5[iii]) was completely unreported (0/72). Only 9 of 74 (12.2%) reports explicitly cited CONSORT-AI. While citing reports showed higher compliance for interaction and data sharing, major technical gaps persisted regardless of citation.

Conclusions: This review provides the first empirical audit distinguishing general nonadherence from structural mismatches where deterministic CONSORT-AI items fail to capture GenAI mechanics. Differing from traditional AI evaluations, it reveals that key reproducibility drivers—system prompts, inference settings, and access modalities—are omitted from current

standards. These empirical findings highlight critical reporting gaps in medical GenAI trials and identify domains for a tailored extension to current reporting standards, although the evidence is limited by the use of nonvalidated signaling questions. Establishing GenAI-specific trial guidelines will ensure that clinical decisions, regulatory approvals, and patient safety safeguards are grounded in reproducible and trustworthy evidence.

Trial Registration: PROSPERO CRD420251034690; <https://www.crd.york.ac.uk/PROSPERO/view/CRD420251034690>

J Med Internet Res 2026;28:e94543; doi: [10.2196/94543](https://doi.org/10.2196/94543)

Keywords: AI; generative artificial intelligence; randomized controlled trials; CONSORT-AI; large language models; reporting adherence; Consolidated Standards of Reporting Trials–AI

Introduction

Generative AI (GenAI), built on foundation models (FMs), is a transformative technology with broad applications, and the medical domain is no exception [1–4]. Unlike conventional AI systems designed for specific purposes, GenAI can perform broader and more complex tasks, offering significant potential for cost-effectiveness [5,6]. Within a few years, these systems have moved from demonstrations of performance on medical knowledge benchmarks [7] to interventions delivered directly to patients and clinicians in clinical settings; as recently as late 2024, a review of registered and published trials of large language models (LLMs) in health care identified only 5 published trials [8]. Realizing such potential in clinical practice demands rigorous validation. Randomized controlled trials (RCTs) remain the gold standard for evaluating the clinical safety and efficacy of any new intervention—a principle that applies equally to GenAI [9].

The results of RCTs serve as foundational evidence for decision-making in policy development, regulation, and clinical practice. To ensure the reliability and transparency of such studies, the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) and the CONSORT (Consolidated Standards of Reporting Trials) guidelines specify the minimum items that must be included in the design and reporting of RCTs [10,11]. In 2020, the CONSORT-AI extension was developed to address potential sources of bias and reporting concerns specific to RCTs involving AI interventions [12]. This extension is used as a benchmark for evaluating reporting in RCTs of AI interventions. The CONSORT-AI development team noted the need to update the guidelines as technology and its applications advance.

Since the publication of CONSORT-AI, its uptake has been assessed in RCTs of AI across health care generally [13,14], in trials appearing after the guideline was issued [15], and within specific fields, including ophthalmology [16] and primary care [17]. These reviews describe a consistent pattern. Aggregate adherence appears moderate to high: median concordance across the checklist was 90% (IQR 77%–94%) among 65 trials [15], and 65% of applicable items were adequately reported among trials in primary care [17]. However, adherence is not evenly distributed across items. This review identified deficiencies in reporting which version of the algorithm was used, how low-quality input data were handled, performance errors and their analysis, and access to the AI intervention or code. In particular, there were no

reported cases of performance errors in the field of ophthalmology [16]. No study to date has provided an in-depth analysis of RCT reporting for GenAI interventions intended for medical purposes. Furthermore, these studies interpret only whether reporting was complied with. That interpretation is sound where a checklist item and the intervention are well matched; however, it does not hold where they have not been examined.

GenAI differs fundamentally from conventional AI in its characteristics and uses [18–21]. The specific differences relevant to reporting can be identified in the reporting guidelines developed for GenAI systems. Transparent reporting of a multivariable model for individual prognosis or diagnosis–LLM (TRIPOD-LLM) targets studies that develop, fine-tune, and evaluate LLMs, requiring the reporting of LLM tuning, outputs, prompt engineering, and updating [21]. Furthermore, the Minimum Information about Clinical AI Checklist for Generative Modeling Research (MI-CLAIM-GEN) was revised to reflect that generative models differ from conventional AI in terms of training, evaluation, interpretability, and reproducibility [22]. Additionally, other frameworks have been proposed to address the provision of health information through chatbots and the development and evaluation of LLMs [23,24]. However, none of these guidelines specifically target RCTs. Given the ongoing evolution and expanding application of AI, the CONSORT-AI extension should also be continuously monitored and reviewed [12].

In this context, identifying and reviewing RCTs of GenAI published to date may help determine opportunities for the broader applicability of the CONSORT-AI extension. Therefore, this systematic review aimed to assess the extent to which RCTs evaluating GenAI for medical purposes adhere to the CONSORT-AI extension checklist items. Through this assessment, we sought to identify gaps between the practical characteristics of health care GenAI interventions and the CONSORT-AI extension, thereby informing future directions for refining GenAI RCT reporting guidelines.

Methods

Overview

This study was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [25–27] (Checklists 1 and 2), and the protocol was prospectively registered in the International

Prospective Register of Systematic Reviews (PROSPERO; CRD420251034690). Methodological changes made after registration, together with their rationale and timing, are summarized in Table S1 in [Multimedia Appendix 1](#). The changes to the eligibility criteria and the search strategy determined the composition of the final sample, and all estimates reported here derive from that sample. The remaining changes affected how adherence was analyzed.

Eligibility Criteria

The inclusion criteria were RCTs involving patients, health care professionals, or laypeople who used GenAI for medical purposes. All randomized designs were eligible, including parallel group and crossover trials. In this study, GenAI was defined as an AI system built on pretrained FMs that produces novel content in response to inputs [28-30]. This definition does not restrict output modalities and encompasses systems that generate text, images, video, and audio [1,28,31]. LLMs represent a subcategory in which the output modality is natural language; systems specialized for clinical applications through fine-tuning, in-context learning, retrieval-augmented generation (RAG), or prompt engineering share the identical generative mechanism and were therefore included [28-30]. Conversely, systems whose outputs are strictly restricted to discriminative results—such as classification, prediction, detection, or risk scoring—even if constructed on pretrained transformer backbones, alongside deterministic, rule-based systems that return prescribed responses, are explicitly excluded [1,29-31]. A medical purpose was defined as an intervention directed at a point in the clinical pathway of an identified recipient. This encompasses interventions concerned with the prevention, screening, diagnosis, monitoring, treatment, or management of disease, injury, or disability, as well as those conveying or interpreting health information in support of such purposes, and those supporting the delivery or documentation of care within a patient encounter. The exclusion criteria were interventions that were not GenAI, interventions not intended for medical purposes, studies that did not involve human participants, studies that were not peer reviewed, conference proceedings, preprints, review articles, non-English publications, and non-RCT designs. Reports that presented only a protocol or a trial registration without results were excluded.

Information Sources

Considering the public release and rapid growth of GenAI since 2022, we searched PubMed, CENTRAL (Cochrane Central Register of Controlled Trials), and Embase for studies published from 2022 through April 3, 2026. All 3 databases were last searched on April 3, 2026. To identify additional eligible trials, backward and forward citation searching was conducted across all initially included studies on August 12, 2026, using the citationchaser (CRAN) tool [32]. To maintain methodological and temporal consistency with the

primary database search, the eligibility of identified records was restricted to studies published or made available online on or before April 3, 2026, while records published after this date were excluded.

Search Strategy

The search used free-text terms and Boolean operators with keywords such as GenAI, LLMs, and RCTs. To identify relevant RCTs, we combined text words and MeSH for GenAI with the Cochrane Highly Sensitive Search Strategy for identifying randomized trials [25]. The detailed search strategy is provided in Table S2 in [Multimedia Appendix 2](#). The search was performed in compliance with PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) [33].

Selection Process

After deduplication, 2 authors (JEK and JWH) independently screened titles and abstracts against the inclusion and exclusion criteria. The 2 authors independently reviewed the full texts to determine final eligibility. Discrepancies were resolved by consensus, and if consensus could not be reached, the disagreements were adjudicated in consultation with a third author (JYK).

Data Collection Process

For the included studies, 2 authors (JEK and JWH) independently extracted data. Bibliographic information (eg, title, authors, year of publication, and journal) and study characteristics (eg, country, sample size, study design, end users of the GenAI, intended use, and outcome measures) were extracted. To identify any information not reported in the main text, published study protocols and trial registrations were also reviewed. In addition, to assess item-level adherence, item-specific signaling questions were developed based on the CONSORT-AI checklist items and the accompanying explanation and elaboration document. Rather than generating new assessment criteria, we extracted every explicit reporting requirement expressed in these source documents, typically phrased as statements such as “should specify,” “should state,” or “should describe,” and converted each requirement into a single closed-ended question answerable from the trial report. The draft questions were developed by one author (JEK) and reviewed against the source documents by a second author (JWH), with discrepancies in wording or in the mapping of a requirement to an item resolved by discussion and adjudicated by a third author (JYK) where necessary. The signaling questions of CONSORT-AI checklist items are listed in [Table 1](#), and the data extraction forms are provided in [Table S3](#) and [Table S4](#) in [Multimedia Appendix 3](#). Disagreements in the assessment of individual reports were resolved by consensus and, when necessary, adjudicated by a third author (JYK).

Table 1. Signaling questions of CONSORT-AI^a checklist items.

Section/topic and numbers	CONSORT-AI checklist	Signaling questions
Title and structured abstract		
1 a, b(i)	Indicate that the intervention involves AI/machine learning in the title and/or abstract and specify the type of model	1. Is it stated in the title and/or abstract that the intervention involves AI/ML^b? 2. Are the terms describing the model type and architecture detailed in the abstract?
1 a, b(ii)	State the intended use of the AI intervention within the trial in the title and/or abstract	1. Is the intended use of the AI intervention stated in the title and/or abstract? 2. Is the disease context described in the title and/or abstract?
Introduction-background and rationale		
2 a(i)	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (for example, health care professionals, patients, and public)	1. Is the intended use of the AI intervention within the clinical pathway included? 2. Is the intended user of the AI intervention included?
Methods-eligible criteria		
4 a(i)	State the inclusion and exclusion criteria at the level of participants	1. Are the participant-level inclusion and exclusion criteria stated?
4 a(ii)	State the inclusion and exclusion criteria at the level of the input data	1. Are the minimum requirements (eg, image resolution, quality metrics, or data format) for the input data specified at the input level? 2. Is it stated when input-data eligibility was assessed? 3. Is it stated how the input-data eligibility was assessed? 4. Is it stated who assessed input-data eligibility?
Methods-trial setting		
4b	Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements	1. Are the hardware and software requirements described (eg, vendor-specific device, computing hardware, cloud integration at each site)? 2. If site-specific changes were required, is the process clearly described (eg, fine-tuning on local data)?
Method-intervention and comparator		
5(i)	State which version of the AI algorithm was used	1. Is the version of the AI system used stated? 2. Is it stated whether the same version was used as in previous supporting studies? 3. Is it stated whether the version changed during the trial? 3-1. If the version changed, are the changes between versions described? 3-2. If the version changed, are the rationales for the changes described? 4. If available, is a regulatory marking reference (eg, a unique device identifier) provided?
5(ii)	Describe how the input data were acquired and selected for the AI intervention	1. Is how the input data were acquired for the AI intervention described? 2. Is how the input data were selected for the AI intervention described? 3. Is how the input data were preprocessed before analysis by the AI system described?
5(iii)	Describe how poor-quality or unavailable input data were assessed and handled	1. Is the amount of poor quality or missing data reported? 2. Is how low-quality or missing data were identified described? 3. Is how low-quality or missing data were handled described? 4. Is any minimum quality standard for the input data stated? 5. If the minimum quality was not met, is it stated where the standard was not achieved? 6. If the minimum quality was not met, is the handling process described including any impact on the participant care pathway? 7. If relevant, is the same information about poor quality or unavailable data reported for the control intervention? 8. If applicable, is it stated whether this minimum quality standard differed from the prerandomization input-data eligibility criteria?
5(iv)	Specify whether there was human-AI interaction in the handling of the input data	1. Is the human-AI interaction stated? 2. Is the human-AI interface used to handle input data described?

Section/topic and numbers	CONSORT-AI checklist	Signaling questions
	and what level of expertise was required of users	3. Is the required user expertise or any user training/instructions for input data handling described?
5(v)	Specify the output of the AI intervention	1. Is the output of the AI intervention clearly specified?
5(vi)	Explain how the AI intervention’s outputs contributed to decision-making or other elements of clinical practice	1. Are downstream interventions that could affect outcomes described? 2. Are the required user expertise and any training or instructions for interpreting the outputs described? 3. If relevant, is comparable information provided for the comparator intervention? 4. Is any discrepancy between intended (as specified in the trial protocol) and actual decision-making reported?
Results-harms		
19	Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, justify why not	1-1. Is it described how performance errors were identified? 1-2. Are the cases of error reported? 1-3. Are the implications of any performance error analysis reported? 1-4. Are the risk-mitigation strategies defined? 2. If no such analysis was planned or done, was the reason explained?
Open science data sharing		
25	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or reuse	1. Is it clearly reported whether and how the AI intervention and/or its code can be accessed or reused?

^aCONSORT-AI: Consolidated Standards of Reporting Trials–AI.
^bML: machine learning.

Extracted text corresponding to the CONSORT-AI checklist was classified into 4 categories. Items were rated as “Reported” only when all underlying signaling questions were met, “Partially reported” when at least one, but not all, signaling questions were satisfied, “Not reported” when none were satisfied, and “Not applicable” when the item could not be applied. Adherence to each checklist item was calculated as the number of reports rated “Reported” divided by the number of reports in which the item was applicable. Ratings of “Not applicable” were assigned when an item was not applicable to the trial design or intervention of a trial and were excluded from the denominator rather than counted as nonadherence. Adherence was presented in tabular form as the number and proportion of trials in each rating category for every checklist item, and graphically as stacked bar charts of the rating distribution. Items were ordered throughout according to their sequence in the CONSORT-AI checklist.

To assess reliability, 2 reviewers independently classified the items, and interrater agreement on these judgments was assessed using Cohen Kappa. To account for potential prevalence paradoxes arising from highly skewed reporting distributions across items, prevalence-adjusted bias-adjusted Kappa (PABAK) values were also calculated. The resultant overall interrater agreement was almost perfect (Kappa=0.81). Item-specific percent agreements, Kappa statistics, and PABAK values are detailed in Table S5 in [Multimedia Appendix 4](#). Concordance was summarized as the adherence rate, defined as the proportion of studies categorized as “Reported” for each CONSORT-AI checklist item.

Synthesis Methods

To evaluate whether CONSORT-AI reporting adherence differed according to study and intervention characteristics,

subgroup analyses were performed across 9 categorical variables: region (North America, Asia, and Europe), publication year (2023, 2024, 2025, and 2026), intended use (information and education, mental health, decision support, documentation, and other), intended user (clinician, laypeople, others, and not reported), AI model family (GPT family, non-GPT, multiple models, and not reported), AI deployment (off-the-shelf, customized, bespoke, not reported), use of vignettes (vignette or nonvignette), reporting of system prompts (system prompt and nonsystem prompt), and declared adherence to CONSORT-AI as a reporting guideline (declared or not declared). These characteristics were selected because they capture study-level context (region and publication year), the clinical setting of the intervention (intended use, intended user, and use of vignettes), and technical attributes of the GenAI intervention itself (model family, deployment, and system prompt), together with declared use of the guideline. For the subgroup analyses, adherence for each CONSORT-AI item was dichotomized into “Reported” versus “Not reported or Partially reported.” The reports in which the item was rated “Not applicable” were excluded from the comparison. To compare adherence rates across subgroups, exact tests were used due to small cell frequencies: the 2-sided Fisher exact test was used for comparisons between 2 groups, and its extension (the Freeman-Halton test) was used for comparisons involving 3 or more groups. Categories that were small or heterogeneous (eg, “Other” in intended use) were excluded from formal significance testing and presented with descriptive statistics. Two-sided $P<.05$ was considered significant. Subgroup comparisons were displayed as a matrix of P values across items and characteristics. Because item-level comparisons were repeated across the checklist items and 9 study characteristics, false-positive findings cannot be excluded.

Benjamini-Hochberg false discovery rate (FDR)–adjusted q values were therefore calculated separately within each of the 9 study characteristics across the applicable items for all comparisons. To address the possibility that items comprising a greater number of signaling questions were less likely to meet the all-or-none criterion for full reporting, adherence was additionally recalculated as the mean proportion of applicable signaling questions satisfied per item as a sensitivity analysis. All statistical analyses were conducted using Python (version 3.10.19; Python Software Foundation).

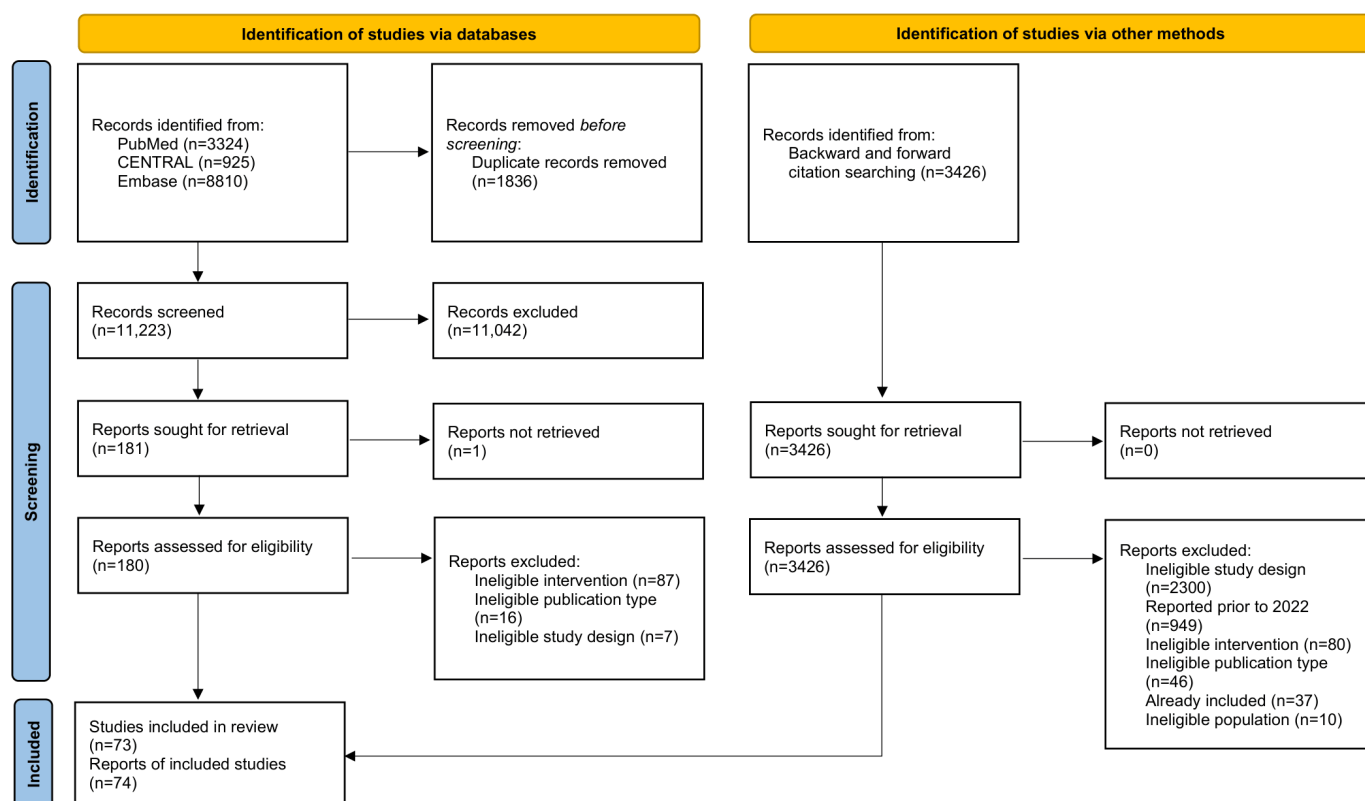
Results

Study Selection

Study selection is presented in [Figure 1](#) according to the PRISMA flow diagram [26]. A total of 13,059 records

were identified from 3 databases. After deduplication, 11,223 records were screened. Following full-text assessment of 181 reports, 111 reports were excluded, including one report whose full text was not accessible despite retrieval attempts [34]. Consequently, a total of 74 reports [35–108] were included in this review. In addition, backward and forward citation searching yielded 3426 records, all of which were sought and assessed for eligibility. Of these, 3422 reports were excluded, resulting in 4 additional eligible reports. In total, 74 published reports representing 73 unique RCTs were included in the review. Two reports [35,36] were separate publications of the same trial and were assessed independently, as the unit of analysis was the individual report.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. The flow diagram outlines the identification, screening, eligibility assessment, and inclusion of trials from database searches and citation searching. A total of 74 published reports [35–108] representing 73 unique RCTs were included in this review.



Study Characteristics

The 74 included reports (representing 73 unique RCTs) were categorized across six domains: (1) geographic region, (2) publication year, (3) primary intended use, (4) intended user, (5) AI model family, and (6) AI deployment mode. Two reports derived from the same trial were evaluated independently at the publication level. In panel (3), “Other” indicates procedural image reconstruction. In panel (5), “Non-GPT” includes non-OpenAI generative models (eg, standalone LLMs, vision-language models, and image reconstruction networks). A total of 74 reports representing 73 RCTs

evaluating GenAI interventions for medical purposes were included in this systematic review, encompassing a cumulative total of 24,150 human participants (median sample size: 100 [IQR 50–247]; range: 9–2671). Notably, all but one [37] of the included reports used LLM-based interventions. The general and technical characteristics of the included reports are detailed in Table S6 in [Multimedia Appendix 5](#).

Geographically, half of the reports were conducted in Asia ($n=37$ [35–38,40,43–45,47–49,51,54,56,58–61,63–66,68,76,77,83,84,86–89,92,93,96,102,104,105], 50.0%), followed by North America ($n=23$ [39,41,42,46,50,52,53,55,57,71,75,78–81,85,91,94,99,103,106–108], 31.1%) and Europe ($n=14$

[62,67,69,70,72-74,82,90,95,97,98,100,101], 18.9%). The publication timeline demonstrated rapid, continuous growth, beginning in 2023 (n=2 [48,49], 2.7%) and 2024 (n=8 [38-41,50-53], 10.8%), reaching a peak in 2025 (n=39 [42-47,54-81,97,99,103,107,108], 52.7%), and continuing into early 2026 (n=25 [35-37,82-96,98,100-102,104-106], 33.8%). Regarding intervention domains, the most common primary intended use was providing medical information and education to laypeople (n=33 [38,44,46,49,51,52,54,57-60,63,64,67-70,72-74,76,83,84,86,89,91-93,96,98,101,102,104], 44.6%), followed by mental health support (n=14 [35,36,47,48,61,65,66,77,78,81,82,87,100,103], 18.9%), clinical decision support (n=14 [39,42,45,53,56,62,71,75,79,80,85,90,94,105], 18.9%), and administrative support/clinical documentation (n=12 [40,41,43,50,55,88,95,97,99,106-108], 16.2%), with one report categorized as other (procedural image reconstruction). Correspondingly, laypeople and patients were the primary end users in 38 reports [35,36,38,40,44,47-49,51-54,58,59,61,63-65,69,70,73,77-79,81-84,86,87,91,93,96,100-104] (51.4%), whereas clinicians were targeted in 23 reports [39,41-43,45,50,55,56,60,62,66,71,75,80,85,92,94,95,97,105-108] (31.1%); end users were unstated or categorized as others in 13 reports [37,46,57,67,68,72,74,76,88-90,98,99] (17.6%).

Technologically, systems built upon the OpenAI GPT model family predominated (n=45 [35,36,39-41,43,47,49-54,57-60,62,64,66-75,77,78,80,81,83,86,88-90,94,95,97-99,102,105], 60.8%), followed by non-GPT (n=22 [37,42,44,45,48,55,56,65,76,79,84,85,87,91,92,100,103,104,106-109], 29.7%) and reports evaluating multiple models simultaneously (n=5 [38,63,82,96,101], 6.8%), the model family was not reported in 2 reports (2.7%). In terms of AI deployment, half of the reports relied on off-the-shelf commercial applications (n=37 [38,39,41,46,49-53,55,57,59,60,62-64,67,68,70-72,74-76,82,89,92,96-99,101,102,105-108], 50.0%), while 32 reports [35,36,40,42-45,47,54,56,58,65,66,69,73,77-81,83-88,90,91,94,95,103,104] (43.2%) used customized architectures (eg, fine-tuned or RAG-integrated systems), and 3 reports [37,48,100] (4.1%) used bespoke models; deployment mode was unreported in 2 reports [61,93] (2.7%). Regarding prompt transparency, only 20 reports [35,40,43-47,62,74,75,78-80,82,83,87,90,94,95,97] (27.0%) explicitly reported the system prompt used in the GenAI intervention. Methodologically, most reports used a parallel-group design (n=67 [35-40,42-44,46-73,75-78,80-94,96,97,100-108], 90.5%), whereas 7 reports [41,45,74,79,95,98,99] (9.5%) used a crossover design. Most reports evaluated 2 parallel/crossover

arms (n=57 [35-38,40,42,44-47,49,51,52,54,56-62,64-69,71,74-81,83-86,89-91,93-95,97-100,102-108], 77.0%), while 17 reports [39,41,43,48,50,53,55,63,70,72,73,82,87,88,92,96,101] (23.0%) featured 3 or more arms. Finally, 17 reports [39,41,42,53,57,62,71,73,75,79,80,94,95,98,99,101,105] (23.0%) used clinical vignettes or standardized scenarios rather than direct real-world patient encounters. Regarding the explicit use of reporting guidelines, a total of 28 reports [35-39,43,46,50,54-56,60-62,65,66,78,80,83,85,88,89,91-93,98,102,105] (37.8%) explicitly cited adopting at least one reporting framework, with some reports following multiple guidelines. Specifically, standard CONSORT was the most frequently cited guideline (n=10 [35,36,39,54,66,78,85,92,102,105], 13.5%), followed by CONSORT-AI (n=9 [37,43,55,56,80,83,88-90], 12.2%), CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Apps and Online Telehealth; n=6 [60-62,65,92,93], 8.1%) and Standards for Quality Improvement Reporting Excellence (SQUIRE; n=3 [38,46,50], 4.1%). Other cited frameworks included TRIPOD-LLM (n=1 [89]), Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health Care (MI-CLEAR-LLM; n=1 [89]), Sex and Gender Equity in Research (SAGER; n=1 [37]), SPIRIT-AI (n=1 [37]), and STROBE (Strengthening the Reporting of Observational Studies in Epidemiology; n=1 [98]). Notably, most reports (n=46 [40-42,44,45,47-49,51-53,57-59,63,64,67-77,79,81,82,84,86,87,90,94-97,99-101,103,104,106-108], 62.2%) did not specify the use of any reporting guideline in their manuscript text.

Results of Syntheses

Reporting Adherence of Included Studies

As shown in Table 2 and Figure 2, item-level adherence to the CONSORT-AI checklist across the 14 evaluated items varied widely among the included reports, ranging from 0% to 98.6% for full reporting. Overall, high adherence was concentrated in general trial reporting items (eg, Items 1 a, b[i], 1 a, b[ii], 2 a[i], and 4 a[i]) and output specifications (Item 5[v]), whereas partial reporting was predominant in trial implementation and technical specifications (eg, Items 4 a[ii], 4b, 5[i], and 5[ii]). Moderate-to-low adherence was observed for clinical decision contribution (Item 5[vi]: 51.5%), data sharing (Item 25: 41.9%), human-AI interaction (Item 5[iv]: 33.8%), and performance error analysis (Item 19: 36.5%). Crucially, reporting of input data quality management (Item 5[iii]) was completely absent across all applicable reports.

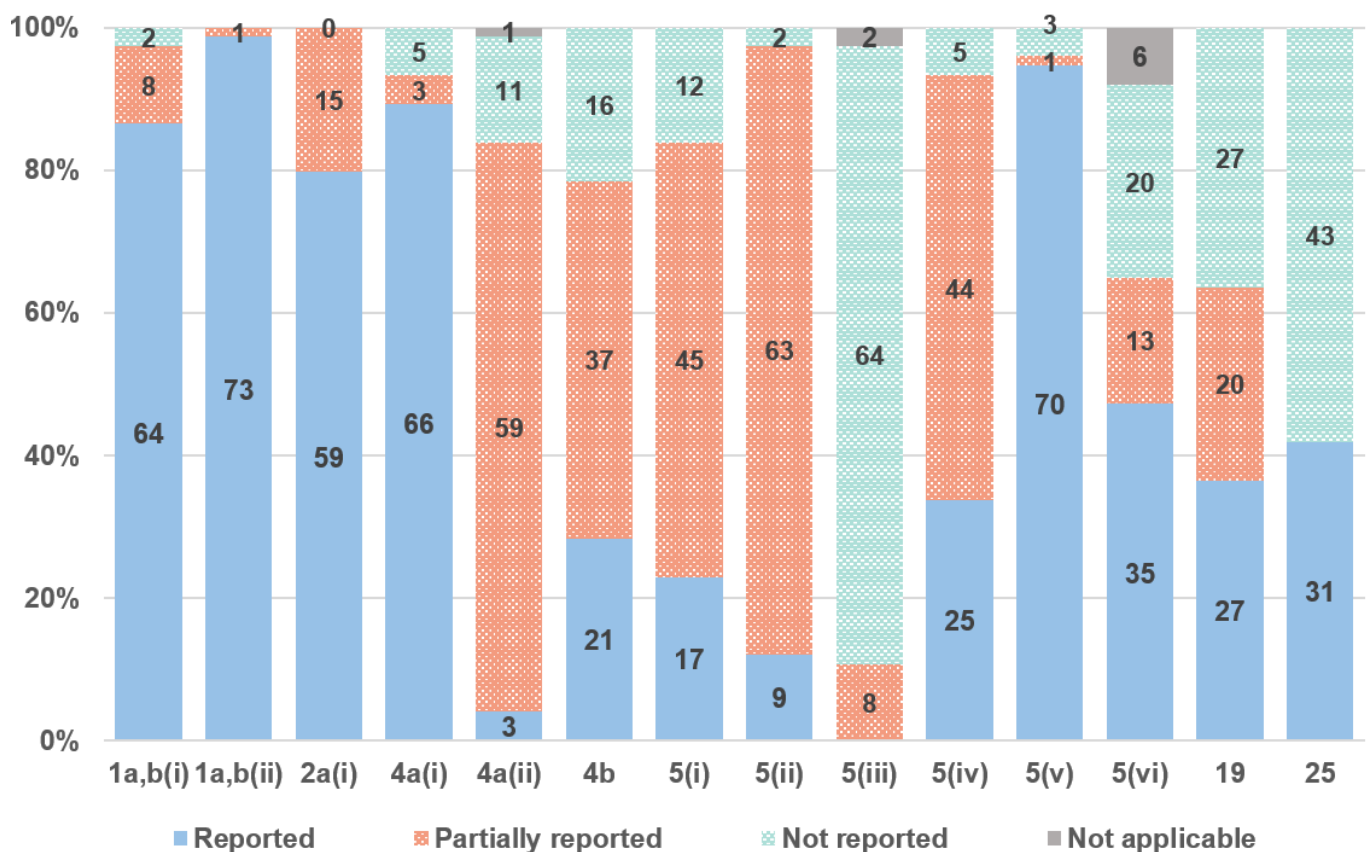
Table 2. Item-level reporting rates of the CONSORT-AI^a checklist.

Section/topic and number	CONSORT-AI checklist	Reporting rate (%)
Title and structured abstract		
1 a, b(i)	Indicate that the intervention involves AI/machine learning in the title and/or abstract and specify the type of model	86.49
1 a, b(ii)	State the intended use of the AI intervention within the trial in the title and/or abstract	98.65
Introduction-background and rationale		

Section/topic and number	CONSORT-AI checklist	Reporting rate (%)
2 a(i)	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (for example, healthcare professionals, patients, and the public)	79.73
Methods-eligible criteria		
4 a(i)	State the inclusion and exclusion criteria at the level of participants	89.19
4 a(ii)	State the inclusion and exclusion criteria at the level of the input data	4.11
Methods-trial setting		
4b	Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements	28.38
Method-intervention and comparator		
5(i)	State which version of the AI algorithm was used	22.97
5(ii)	Describe how the input data were acquired and selected for the AI intervention	12.16
5(iii)	Describe how poor-quality or unavailable input data were assessed and handled	0.00
5(iv)	Specify whether there was human–AI interaction in the handling of the input data and what level of expertise was required of users	33.78
5(v)	Specify the output of the AI intervention	94.59
5(vi)	Explain how the AI intervention’s outputs contributed to decision-making or other elements of clinical practice	51.47
Results-harms		
19	Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, justify why not	36.49
Open science data sharing		
25	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or reuse	41.89

^aCONSORT-AI: Consolidated Standards of Reporting Trials–AI.

Figure 2. Item-level reporting status of CONSORT-AI (Consolidated Standards of Reporting Trials–AI) checklist. This stacked bar chart shows the number of studies that fully reported, not reported, partially reported, or not applicable for each CONSORT-AI checklist item. Each bar represents one checklist item, with color segments indicating reporting adherence.



Ratings of “Not applicable” accounted for 0.9% (9/1036) of the item-level observations and were confined to 3 items: Item 5(vi; n=6), Item 5(iii; n=2), and Item 4 a(ii; n=1). Denominators were therefore 74 for 11 items, 73 for Item 4 a(ii), 72 for Item 5(iii), and 68 for Item 5(vi). In a sensitivity analysis, recalculating adherence as the mean proportion of applicable signaling questions satisfied per report did not alter the overall pattern of findings, although estimates were higher for Items 5(ii; 46.2%, SD 24.4%) and 5(iv; 68.7%, SD 28.8%), indicating that partial reporting was common for these items. Item 5(iii) remained near-absent (5.3%, SD 15.9%; Table S7 in [Multimedia Appendix 6](#)).

Subgroup Analyses

Item-level adherence was compared across subgroups defined by region, publication year, intended use, intended user, AI model family, AI deployment, use of clinical vignettes, use of a system prompt, and whether the report cited CONSORT-AI as a reporting guideline. Item-level *P* values for all comparisons are presented in [Figure 3](#), and group-specific proportions for each variable are provided in Tables S8-S16 in [Multimedia Appendix 7](#). Item 5(iii) was not testable for any variable, as no article reported it.

Figure 3. Item-level *P* values for subgroup comparisons of CONSORT-AI (Consolidated Standards of Reporting Trials–AI) adherence (n=74). Each cell shows the *P* value from an exact test comparing the proportion of reports in which the item was classified as “Reported” across groups of the corresponding variable. For cells with unadjusted *P*<.05, the FDR-adjusted *q* value is also displayed. Dark pink cells indicate differences retained after adjustment; light pink cells indicate differences not retained after adjustment. Item 5(iii) was not testable because no trial reported it. The group-specific proportions and Benjamini-Hochberg adjusted *q* values are provided in Table S8-S16 in [Multimedia Appendix 7](#). CONSORT-AI: Consolidated Standards of Reporting Trials–AI.

	Region	Publication year	Intended use	Intended user	AI model family	AI deployment	Clinical vignette	System prompt	CONSORT-AI declared
1a,b(i) AI intervention stated in title/abstract	P=.03 q=.20	P=.35	P<.001 q<.001	P=.40	P=.78	P=.05	P=.10	P≥.99	P≥.99
1a,b(ii) Intended use stated in title/abstract	P=.50	P≥.99	P=.55	P≥.99	P≥.99	P=.49	P≥.99	P=.27	P≥.99
2a(i) Intended use and user in clinical pathway	P=.59	P=.67	P=.04 q=.13	P<.001 q<.001	P=.48	P=.01 q=.06	P=.74	P=.75	P≥.99
4a(i) Participant-level eligibility criteria	P=.007 q=.09	P=.90	P=.19	P=.06	P=.63	P≥.99	P<.001 q<.001	P≥.99	P=.58
4a(ii) Input-data-level eligibility criteria	P=.17	P≥.99	P≥.99	P=.36	P≥.99	P=.65	P≥.99	P=.02 q=.08	P=.33
4b Integration into the trial setting	P=.44	P=.09	P=.17	P=.63	P=.01 q=.17	P<.001 q=.01	P=.13	P=.15	P=.71
5(i) Version of the AI algorithm	P=.58	P=.82	P=.25	P=.80	P≥.99	P=.75	P=.75	P=.21	P=.20
5(ii) Input-data acquisition and selection	P=.45	P=.14	P=.63	P=.11	P=.63	P≥.99	P≥.99	P=.001 q=.01	P=.30
5(iii) Handling of poor-quality or missing data	not testable	not testable	not testable	not testable	not testable	not testable	not testable	not testable	not testable
5(iv) Human–AI interaction and user expertise	P=.58	P=.22	P=.04 q=.13	P=.18	P=.38	P=.32	P≥.99	P≥.99	P=.006 q=.04
5(v) Output of the AI intervention	P=.70	P=.81	P=.39	P=.08	P≥.99	P=.26	P=.22	P=.57	P≥.99
5(vi) Contribution of output to decision-making	P=.15	P=.29	P<.001 q<.001	P<.001 q<.001	P=.05	P=.03 q=.10	P=.03 q=.23	P=.59	P=.03 q=.12
19 Performance errors and error analysis	P=.56	P=.75	P=.66	P=.09	P=.54	P=.50	P=.78	P=.42	P=.27
25 Access to the AI intervention or code	P=.52	P≥.99	P=.06	P=.56	P=.44	P=.03 q=.10	P≥.99	P=.02 q=.08	P=.003 q=.04

■ *P*<.05, retained after adjustment (*q*<.05)

■ *P*<.05, not retained after adjustment

■ *P*≥.05

The region showed differences in Items 1 a, b(i) and 4 a(i), though neither retained significance after FDR adjustment, and no Item differed by publication year. Intended use showed differences in Items 1 a, b(i), 2 a(i), and 5(vi), although only the differences in Items 1 a, b(i) and 5(vi) retained significance after adjustment, with 5(vi) separating reports in clinical decision support and clinical documentation from those in patient information and mental health. Differences by intended user were retained after adjustment for Items 2 a(i) and 5(vi). That for 5(vi) was the largest difference observed in any comparison in this review. Every clinician-facing trial reported how the intervention’s output contributed to decision-making, whereas fewer than one in 5 layperson-facing reports did so. The difference in Item 2 a(i) was definitional rather than substantive. The reports

were classified as “not reported” because they did not state an intended user, as required by this item. An unadjusted difference by AI model family in Item 4b did not withstand FDR adjustment. AI deployment showed differences in Items 2 a(i), 4b, 5(vi), and 25, though only Item 4b retained significance after adjustment. The use of clinical vignettes showed a difference in Item 4 a(i), which was retained after adjustment, whereas the use of a system prompt significantly distinguished the input-related Item 5(ii). Nine reports [37,43,55,56,80,83,88–90] declared adherence to CONSORT-AI as a reporting guideline. Across 14 items, these articles reported 64.0% (80/125) of applicable items, higher than 46.6% (420/902) among the remaining 65 reports [35,36,38–42,44–54,57–79,81,82,84–87,91–108] (denominators exclude 1 and 8 item-level ratings of “Not applicable,” respectively), with

differences retained after adjustment in Items 5(iv) and 25. Four items showed no difference across any of the 9 variables examined: 1 a, b(ii), 5(i), 5(v), and 19.

Discussion

Principal Findings

This systematic review evaluated the reporting adherence of 74 published reports representing 73 RCTs evaluating medical GenAI interventions against the CONSORT-AI checklist [12]. This discussion first examines the critical reporting gaps identified across the included trials and outlines key considerations for refining future reporting standards tailored to GenAI-based medical interventions. Our evaluation revealed substantial reporting deficiencies across critical methodological and technical domains, including complete omissions in input data quality management (5[iii]), frequent partial reporting of algorithm versioning (5[i]), and insufficient descriptions of human–AI interaction (5[iv]) and performance error analyses (19). In parallel with these widespread reporting gaps, explicit adoption of the CONSORT-AI guideline remained low, with only 9 [37,43,55,56,80,83,88–90] (12.2%) of included reports citing its use.

These observations invite different explanations. An item may go unreported because investigators were unaware that it was required, in which case the deficiency lies with adherence, and wider dissemination of the guideline is the remedy. It may go unreported because what the item asks for does not correspond to how a GenAI intervention is specified or delivered, in which case the deficiency lies with the item and the remedy is revision. Or the information needed to reproduce the report may not be requested by any item at all, in which case neither adherence nor revision suffices and the checklist must be extended.

Studies that explicitly declared adherence to CONSORT-AI demonstrated higher compliance for code/model accessibility (Item 25) and input interaction requirements (Item 5[iv]) compared to nonciting studies ($P<.05$; $q<.05$). While an unadjusted difference was also observed for output decision-making contribution (Item 5[vi]), this association was attenuated after FDR adjustment and represents an exploratory finding. These statistical findings suggest that reporting opacities in Items 5(iv) and 25 are attributable to investigator nonadherence, which can be mitigated by enforcing existing reporting standards.

Specifying the instructions or knowledge required for users to interpret outputs and detect potential errors is needed, yet reporting for this domain remained limited (Item 5[vi]: 51.5%). This gap is further reflected in our subgroup analysis: reports with clinicians as the intended AI users achieved a 100% (23/23) full reporting rate for Item 5(vi), whereas reports with laypeople as the intended AI users achieved a full reporting rate of 15.8% (6/38; $P<.001$). Laypeople bring no baseline expertise, and the interpretive burden that GenAI transfers to the user is correspondingly heavier [110]. Reports

of laypeople-facing interventions should therefore state what instructions participants were given for interpreting outputs.

Additionally, the level of professional expertise or guidance required during human–AI interaction while handling input data (Item 5[iv]) was fully reported in only 33.8% of reports. Reporting varied across clinical application domains: it was higher in domains where clinicians are the primary intended users—such as clinical decision support (6/14, 42.9%) and administrative/clinical documentation (7/12, 58.3%)—and lower in domains where laypeople are the primary intended users, such as informational and educational support (10/33, 30.3%) and mental health support (1/14, 7.1%). Although this domain-level variation was exploratory after FDR adjustment, this pattern suggests that trials of laypeople-facing conversational tools less often report the user instructions and interaction safeguards required for operating GenAI interventions.

Beyond user interaction, reporting regarding open science principles was similarly suboptimal, with intervention and code accessibility (Item 25) fully specified in only 41.9% of reports. Importantly, without a statement of whether and how the intervention or its underlying prompts, parameters, and code can be obtained, subsequent researchers cannot access the exact intervention evaluated, let alone validate it [111]. Although subgroup differences across deployment modes did not retain statistical significance after adjustment, this low reporting was consistent with trials using commercial off-the-shelf models, where the artifact is controlled by a third party and may not persist in the form tested [112].

By contrast, explicit citation of CONSORT-AI was not associated with higher reporting adherence for the other items. This pattern suggests that while enforcing CONSORT-AI can resolve deficiencies in general interaction and accessibility items (Items 5[iv] and 25), it remains insufficient on its own to rectify reporting gaps in remaining deployment and technical specifications (such as Items 4b, 4 a[ii], 5[i], 5[ii], and 5[iii]). The remainder of this discussion therefore examines each of these items in turn—how it was drafted, why it fails for GenAI interventions, and what form it would need to take instead—before considering the conditions that determine GenAI output but that no CONSORT-AI item presently requests at all.

Descriptions of how the GenAI intervention was integrated into the trial environment, including site-specific or external requirements (Item 4b), were reported in 28.4% of reports. Given that most included interventions were LLMs or commercial API-based services, the remaining studies did not report baseline operational prerequisites, such as required internet connectivity or target access devices (eg, mobile phones or desktop computers). Subgroup analysis showed that this reporting rate varied by the deployment mode. Reports using commercial off-the-shelf models (eg, ChatGPT) demonstrated a significantly lower full reporting rate for Item 4b (5/37, 13.5%) compared to customized (12/32, 37.5%) or bespoke (3/3, 100%) interventions ($P=.001$; $q=.012$). Conversely, while reports using GPT family models exhibited lower reporting adherence for Item 4b than

non-GPT models ($P=.01$), this difference was attenuated after adjustment ($q=.173$), indicating that deployment rather than AI model family drives reporting transparency. This disparity highlights the possibility that investigators using ubiquitous commercial models frequently operate under the assumption that widespread familiarity obviates the need for explicit documentation of basic connectivity and deployment prerequisites.

The 3 CONSORT-AI items concerning input data had the lowest reporting rates in this review. Input-data eligibility criteria 4 a(ii) was reported in 4.1% of reports, input-data acquisition and selection (5(ii)) were reported in 12.2%, and handling of poor quality or unavailable input data 5(iii) was reported by none. All 3 items were drafted for interventions that operate on data assembled before a trial begins, whereas, in 73 of the 74 reports, we reviewed an LLM-received text generated during the trial itself. In conventional AI systems, input data such as medical images or biosignals are structured and quantifiable, allowing researchers to assess their quality in terms of resolution, data format, and quality metrics [12]. Since these inputs can be standardized through the preprocessing procedure, Items 4 a(ii), 5(ii), and 5(iii) of the CONSORT-AI could be systematically reported. In contrast, in most GenAI systems reported here, the input is composed at the point of use, most often by the intended user, and its properties are determined by that person rather than by the protocol. The same intervention consequently receives materially different input from one participant to the next. This variation is not incidental to performance. Language models have been shown to vary in accuracy by 8%-50% according to how a query is optimized [113] and by 5%-18% according to the order in which the same information is presented [114].

Item 4 a(ii) of the CONSORT-AI checklist requires researchers to specify the minimum input requirements necessary for the AI intervention to achieve its intended use prior to randomization. This item is essential for ensuring reproducibility under consistent conditions and enhancing the reliability of result interpretation. As noted earlier, it is inherently difficult to manage inputs in clinical trials involving GenAI, which likely contributes to the low level of adherence to this reporting item. Nevertheless, it is necessary to predefine at least basic inclusion and exclusion criteria for inputs at the prompt level. Allowing users to freely generate prompts can be beneficial for assessing real-world applicability and understanding how input variation affects output quality. However, the absence of predefined input criteria undermines reproducibility [115]. Item 4 a(ii) is therefore best revised rather than removed. Where investigators structure the input—through prompt templates, standardized scenarios, or predefined instructions—eligibility can be stated at the level of that material: the language in which input is accepted, permissible length, the clinical elements it must contain for the intervention to function as intended, and the categories of content that must not be submitted. Where unconstrained interaction is intentional to assess real-world performance, that decision should be stated explicitly and its consequences acknowledged.

Item 5(ii), which asks how input data were acquired and selected for the AI intervention prior to analysis by the AI system, was fully reported in only 9 reports [43,45,46,60,62,74,78,90,99] (12.2%) and by only one [78] of the 38 reports [35,36,38,40,44,47-49,51-54,58,59,61,63-65,69,70,73,77-79,81-84,86,87,91,93,96,100-104] in which laypeople were the intended users. In our evaluation, full reporting was credited exclusively to reports that defined explicit data gating procedures. These included either structural constraints at the user interface level—such as enforcing structured input templates or predefined prompt instructions—or human-in-the-loop protocols in which clinical implementers or investigators filtered, verified, and submitted patient data into the AI system on the users' behalf. Conversely, trials permitting unconstrained, free-text generation by interacting users were categorized as nonreporting because the tasks of data selection, structuring, and preprocessing were entirely delegated to the individual intended user without investigator oversight.

This empirical finding highlights a major methodological vulnerability in conversational GenAI trials: when user interaction is unconstrained, investigators may treat input data handling not as a controlled methodological step but as an unreportable artifact of user behavior. However, unconditioned user inputs introduce variability that directly impacts GenAI performance [38,39,116]. Even in unconstrained, free-text conversational interventions, investigators control input selection and acquisition at the backend through prompt engineering architectures, such as system prompts, or automated input-filtering rules prior to model inference [40-45]. This is directly shown by our subgroup analysis, where reporting of a system prompt was associated with higher adherence to Item 5(ii) after adjustment ($P=.001$; $q=.014$).

Therefore, our findings demonstrate that Item 5(ii) remains highly relevant but requires explicit contextual adaptation for GenAI. Future reporting standards should refine Item 5(ii) to mandate disclosure at 2 operational levels: first, at the user interface level, by stating whether inputs are governed by structured templates, clinician mediation, or unconstrained free text; and second, at the system architecture level, by detailing backend prompt engineering, context selection, and preinference filtering protocols executed before the AI system analyzes the query.

Item 5(iii) was reported by no trial in this review. Reviews of conventional AI trials have also found this item to be among the least reported [13,14,16,17], and to that extent the deficiency is not new. Its character in GenAI trials is nonetheless different and, in our view, more consequential. Item 5(iii) requires authors to describe how low-quality or missing input data are handled. For structured input, "poor quality" has a referent that can be defined in advance and detected automatically; for example, an image below a resolution threshold, a corrupted file, or a missing field. When user queries serve as inputs, ambiguity, inconsistency, lack of context, or omission of key terms can degrade input quality [117,118]. The difficulty compounds where the intervention involves sustained interaction. As the interaction continues,

each turn generates new input that cannot be screened before it reaches the model.

In this context, unavailable data can represent inputs that should not be submitted, namely protected health information (PHI). PHI includes identifiers that can be used to recognize an individual and cover past, present, or future physical or mental health conditions [119]. In the United States, PHI is strictly regulated under the Health Insurance Portability and Accountability Act (HIPAA). Any GenAI system handling PHI must comply with HIPAA requirements like any other health care technology. However, the web-based interface of ChatGPT, the most used model in the reviewed studies, is not HIPAA-compliant [120]. This means patient information must never be entered into the public web interface. Instead, a HIPAA-eligible cloud environment that offers a Business Associate Agreement (BAA), such as the Azure OpenAI Service [121], should be used. The trials in this review were conducted in 16 countries, with 21 reports [39,41,42,46,50,53,55,57,71,75,78-81,85,91,94,99,103,107,108] from the United States. Trials conducted elsewhere are governed by their own regimes, including the General Data Protection Regulation (GDPR) in Europe [122], and what constitutes adequate protection is therefore not a single standard. What is common across them is that the answer depends on the interface through which data were submitted and the contractual arrangement under which they were processed, which is precisely what trial reports do not currently state. Eight reports addressed such matters partially and none completely.

Trial reports should state what constitutes an inadmissible input, what interface arrangement governed the data submitted, what measures prevented participants or operators from submitting inadmissible content, and what was done when this occurred. MI-CLAIM-GEN takes a comparable approach, asking not how such data were processed but what steps were taken to understand privacy and security concerns and distinguishing between discussion, identification, and mitigation of them [22].

Item 5(i) was reported in 23.0% of articles, and adherence did not differ across any of the 9 variables examined. Version reporting has been found deficient in reviews of conventional AI trial reports [13-16], but the 2 settings differ in whether the intervention stays fixed once the trial begins. A conventional model evaluated in a trial is ordinarily locked, and its identity is stable for as long as the trial runs. Off-the-shelf GenAI systems are not. They are continuously updated with new training data, model weights, and fine-tuning methods [123,124]. Many of these models are updated on short cycles and rolled out without public changelogs, making it difficult to ensure consistent study outcomes [125]. Even models under the same name, such as GPT-4, exist in multiple versions (eg, gpt-4-0314, gpt-4-0613, gpt-4-turbo, and gpt-4o), each with distinct performance profiles [125]. An article that reports having used “GPT-4” has partially satisfied Item 5(i) as written but has not supplied what the item exists to secure. The difficulty is not only that version labels are coarse. Item 5(i) refers to the version of “the AI algorithm,” in the singular, whereas a contemporary GenAI intervention

is frequently an assembly of components. MI-CLAIM-GEN accordingly asks for the base model together with any embedding models, retrieval models, and auxiliary models or tools [22], and TRIPOD-LLM asks for the model name, version, and last date of training, and separately for the architecture, fine-tuning procedure, and alignment strategy [21]. An intervention combining a base model with retrieval over a local document set and a fixed set of instructions cannot be identified by a version string, however precise. That an item can be met without meeting its purpose is a limitation of the item rather than of the authors who complied with it, and it indicates that the reproducibility requirement is incompletely specified. Reproducibility is further complicated by the method of access. Interacting with an LLM through a web interface can yield different results than using an API, as the former may include hidden system prompts or integrated tools like web browsing. In contrast, API access provides a more controlled, researcher-defined environment. These differences can significantly alter a model’s factual accuracy, performance, and reasoning transparency [109,126]. Despite these critical variables, many studies fail to report the specific model version, offering only generic descriptions that limit transparency and hinder replication. Therefore, to ensure transparency and reproducibility in GenAI-based clinical trials, reports should specify the exact model version, the date of use, and the access method (eg, API or web interface) [127].

Two further conditions determine what a GenAI intervention produces, and neither is requested by any CONSORT-AI item. The system prompt establishes the conditions under which a model generates output, and it can shift reading level, enforce consistency across participants, and suppress content the investigators do not intend the intervention to produce [43,46,47]. The inference settings determine how far output may vary under those terms. Their influence is comparable to that of the model version itself. In clinical summarization, adding instruction prompts improves performance but also increases performance variance, so that prompts sharing similar meanings yield markedly different summaries [128]. In a simulation study in which 4 emergency medicine cases were each submitted to GPT-4o 250 times at 5 temperature settings, leading diagnostic accuracy with physical examination data fell from 100% at a temperature of 0.0% to 89.4% at 1.0, while the mean number of distinct diagnoses generated across repetitions rose from 4.5 to 26.25 [129]. Two trials evaluating the same model at different inference settings and system prompts are not evaluating the same intervention at all.

Taken together, these conditions define what would have to be reported for a GenAI intervention to be reproduced: the base model and any auxiliary components, the version and the date on which it was used, the method of access, the system prompt, and the inference parameters under which output was generated. Every one of these is requested by at least one of the 2 reporting frameworks related to GenAI [21,22]. Of these elements, CONSORT-AI presently requests one model version. A trial article reporting only the model version has satisfied Item 5(i) while leaving the

intervention unidentifiable. Where an item can be met without its reproducibility being served, the requirement is incompletely specified rather than poorly reported, and revision for GenAI trials merits consideration. Table S17 in [Multimedia Appendix 8](#) consolidates the challenges identified throughout this discussion, distinguishing those arising from items whose underlying construct does not fit a generative intervention from those for which no corresponding item exists.

Almost every element our analysis identifies as missing or unaskable is already established elsewhere: TRIPOD-LLM [21] specifies model name, version, training date, architecture, fine-tuning, and alignment; MI-CLAIM-GEN [22] specifies base and auxiliary components together with structured privacy and security reporting. However, neither framework was developed specifically for randomized trials. These findings identify candidate domains for an extension of CONSORT-AI that would import these settled requirements into the trial setting, restating the input-data items at both the interface and backend levels, replacing the single algorithm version with a full intervention specification, requiring that inadmissible input and the contractual arrangement governing it be defined in advance under national regimes, and naming GenAI-specific error classes under Item 19. So defined, the extension would make reproducible reporting achievable for the class of intervention that now accounts for the great majority of GenAI trials, without duplicating what the wider reporting ecosystem has already settled.

Several limitations of this study should be noted. First, no validated instrument exists for item-level CONSORT-AI assessment, so we developed signaling questions ourselves. Although each question was derived directly from an explicit reporting requirement stated in the CONSORT-AI statement, they have not been externally validated, and a different operationalization could yield different adherence estimates. In particular, requiring all signaling questions to be satisfied for an item to be classified as fully reported is a conservative rule, and items comprising a greater number of requirements are correspondingly less likely to meet it. That said, all assessments were performed independently in

duplicate, and item-specific agreement statistics are reported in Table S5 in [Multimedia Appendix 4](#). Second, we restricted inclusion to peer-reviewed, English-language studies explicitly identified as RCTs. This preserved methodological homogeneity but may have excluded randomized evaluations described using other terminology, published in other languages, or available only as preprints, and the excluded studies may differ systematically in reporting practice. Finally, the trials reviewed were predominantly evaluations of text-based language models. As multimodal interventions increasingly enter clinical trials, they will inevitably raise new methodological and reporting challenges that the present CONSORT-AI checklist does not anticipate. To support the continuous updating of these reporting standards, we have prospectively identified and will continue to monitor relevant preprints, trial registrations, and conference abstracts.

Conclusion

RCTs of GenAI for medical purposes adhere unevenly to CONSORT-AI. Trial metadata and output specification were reported by most trial reports, whereas input data, technical execution, and reproducibility items were reported by few, and handling of poor-quality or missing input data was not reported by any trial. These deficiencies are of 2 kinds. Some reflect nonadherence to requirements that already exist and could be addressed by wider dissemination of the current guideline, while others reflect a mismatch between what the checklist asks and how a generative intervention is specified and delivered and cannot be resolved by adherence alone. Development of a formal reporting extension would require a structured consensus process, as recommended for the development of health research reporting guidelines [130] and as followed in the development of CONSORT-AI itself [12], involving trialists, clinicians, regulators, AI developers, and reporting guideline methodologists. This study establishes that basis, providing the first item-level account of how GenAI trials report against CONSORT-AI and suggesting what such an extension would have to specify to make these trials reproducible and interpretable.

Generative AI Use

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing, adapting and adjusting emotional tone, translation. The GenAI tool used was: Claude Opus 5 (Anthropic), Gemini 3.6 Flash (Google). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by: JEK, JWH, SMK, and JYK.

Funding

This research was supported by a grant RS-2024-00331775 from Ministry of Food and Drug Safety in 2026 and this research was supported by a grant RS-2026-25511885 from Ministry of Food and Drug Safety in 2026. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Data Availability

Data collected and used in this systematic review can be requested from corresponding author.

Authors' Contributions

Conceptualization: JEK (equal), JWH (equal)

Data curation: JEK (equal), JWH (equal)

Formal analysis: JEK (equal), JWH (equal)

Funding acquisition: SMK

Investigation: JEK (equal), JWH (equal)

Methodology: JEK (equal), JWH (equal)

Project administration: JYK

Supervision: JYK (lead), SMK (supporting)

Visualization: JEK (equal), JWH (equal)

Writing – original draft: JEK (equal), JWH (equal)

Writing – review & editing: JEK (equal), JWH (equal), SMK (supporting), JYK (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1

S1 Comparison of registered protocol.

[\[DOCX File \(Microsoft Word File\), 20 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Search strategy.

[\[DOCX File \(Microsoft Word File\), 17 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Data extraction form.

[\[DOCX File \(Microsoft Word File\), 35 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

S5 item-specific agreement statistics.

[\[DOCX File \(Microsoft Word File\), 21 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Characteristics of the included reports.

[\[DOCX File \(Microsoft Word File\), 61 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Signaling question sensitivity analysis.

[\[DOCX File \(Microsoft Word File\), 18 KB-Multimedia Appendix 6\]](#)

Multimedia Appendix 7

Subgroup analysis.

[\[DOCX File \(Microsoft Word File\), 49 KB-Multimedia Appendix 7\]](#)

Multimedia Appendix 8

S17 Characteristics of Gen AI interventions identified in this review.

[\[DOCX File \(Microsoft Word File\), 17 KB-Multimedia Appendix 8\]](#)

Checklist 1

PRISMA 2020 checklist.

[\[PDF File \(Adobe File\), 343 KB-Checklist 1\]](#)

Checklist 2

SWiM checklist.

[\[PDF File \(Adobe File\), 302 KB-Checklist 2\]](#)

References

1. Banh L, Strobel G. Generative artificial intelligence. Electron Markets. Dec 2023;33(1). [doi: [10.1007/s12525-023-00680-1](https://doi.org/10.1007/s12525-023-00680-1)]
2. Dogru T (Dr. True), Line N, Mody M, et al. Generative artificial intelligence in the hospitality and tourism industry: developing a framework for future research. J Hosp Tour Res. Feb 2025;49(2):235-253. [doi: [10.1177/10963480231188663](https://doi.org/10.1177/10963480231188663)]

3. McDonald N, Johri A, Ali A, Collier AH. Generative artificial intelligence in higher education: evidence from an analysis of institutional policies and guidelines. *Comput Hum Behav Artif Hum*. Mar 2025;3:100121. [doi: [10.1016/j.chbah.2025.100121](https://doi.org/10.1016/j.chbah.2025.100121)]
4. Shahriar S, Corradini MG, Sharif S, Moussa M, Dara R. The role of generative artificial intelligence in digital agri-food. *J Agric Food Res*. Apr 2025;20:101787. [doi: [10.1016/j.jafr.2025.101787](https://doi.org/10.1016/j.jafr.2025.101787)]
5. Goyal A, Tariq MD, Ahsan A, Brateanu A. Can ChatGPT 4.0 diagnose acute aortic dissection? Integrating artificial intelligence into medical diagnostics. *Am J Cardiol*. Mar 15, 2025;239:90-92. [doi: [10.1016/j.amjcard.2025.01.010](https://doi.org/10.1016/j.amjcard.2025.01.010)] [Medline: [39826879](https://pubmed.ncbi.nlm.nih.gov/39826879/)]
6. Roberts LJ, Jayasena R, Khanna S, Arnott L, Lane P, Bain C. Challenges for implementing generative artificial intelligence (GenAI) into clinical healthcare. *Intern Med J*. Jul 2025;55(7):1063-1069. [doi: [10.1111/imj.70035](https://doi.org/10.1111/imj.70035)] [Medline: [40135733](https://pubmed.ncbi.nlm.nih.gov/40135733/)]
7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. Aug 2023;620(7972):172-180. [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
8. Omar M, Nadkarni GN, Klang E, Glicksberg BS. Large language models in medicine: a review of current clinical trials across healthcare applications. *PLoS Digit Health*. Nov 2024;3(11):e0000662. [doi: [10.1371/journal.pdig.0000662](https://doi.org/10.1371/journal.pdig.0000662)] [Medline: [39561120](https://pubmed.ncbi.nlm.nih.gov/39561120/)]
9. Kohane IS. Injecting artificial intelligence into medicine. *NEJM AI*. Jan 2024;1(1):AIe2300197. [doi: [10.1056/AIe2300197](https://doi.org/10.1056/AIe2300197)]
10. Hróbjartsson A, Boutron I, Hopewell S, et al. SPIRIT 2025 explanation and elaboration: updated guideline for protocols of randomised trials. *BMJ*. Apr 28, 2025;389:e081660. [doi: [10.1136/bmj-2024-081660](https://doi.org/10.1136/bmj-2024-081660)] [Medline: [40294956](https://pubmed.ncbi.nlm.nih.gov/40294956/)]
11. Hopewell S, Chan AW, Collins GS, et al. CONSORT 2025 statement: updated guideline for reporting randomized trials. *Nat Med*. Jun 2025;31(6):1776-1783. [doi: [10.1038/s41591-025-03635-5](https://doi.org/10.1038/s41591-025-03635-5)] [Medline: [40229553](https://pubmed.ncbi.nlm.nih.gov/40229553/)]
12. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. Sep 2020;26(9):1364-1374. [doi: [10.1038/s41591-020-1034-x](https://doi.org/10.1038/s41591-020-1034-x)] [Medline: [32908283](https://pubmed.ncbi.nlm.nih.gov/32908283/)]
13. Shahzad R, Ayub B, Siddiqui MAR. Quality of reporting of randomised controlled trials of artificial intelligence in healthcare: a systematic review. *BMJ Open*. Sep 5, 2022;12(9):e061519. [doi: [10.1136/bmjopen-2022-061519](https://doi.org/10.1136/bmjopen-2022-061519)] [Medline: [36691151](https://pubmed.ncbi.nlm.nih.gov/36691151/)]
14. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw Open*. Sep 1, 2022;5(9):e2233946. [doi: [10.1001/jamanetworkopen.2022.33946](https://doi.org/10.1001/jamanetworkopen.2022.33946)] [Medline: [36173632](https://pubmed.ncbi.nlm.nih.gov/36173632/)]
15. Martindale APL, Llewellyn CD, de Visser RO, et al. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines. *Nat Commun*. Feb 2024;15(1):38388497. [doi: [10.1038/s41467-024-45355-3](https://doi.org/10.1038/s41467-024-45355-3)]
16. Pattathil N, Zhao JZL, Sam-Oyerinde O, Felfeli T. Adherence of randomised controlled trials using artificial intelligence in ophthalmology to CONSORT-AI guidelines: a systematic review and critical appraisal. *BMJ Health Care Inform*. Jul 2023;30(1):e100757. [doi: [10.1136/bmjhci-2023-100757](https://doi.org/10.1136/bmjhci-2023-100757)] [Medline: [37463773](https://pubmed.ncbi.nlm.nih.gov/37463773/)]
17. Zhong J, Zhu T, Huang Y. Reporting quality of AI intervention in randomized controlled trials in primary care: systematic review and meta-epidemiological study. *J Med Internet Res*. Feb 25, 2025;27:e56774. [doi: [10.2196/56774](https://doi.org/10.2196/56774)] [Medline: [39998876](https://pubmed.ncbi.nlm.nih.gov/39998876/)]
18. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. Apr 2023;616(7956):259-265. [doi: [10.1038/s41586-023-05881-4](https://doi.org/10.1038/s41586-023-05881-4)] [Medline: [37045921](https://pubmed.ncbi.nlm.nih.gov/37045921/)]
19. Howell MD. Generative artificial intelligence, patient safety and healthcare quality: a review. *BMJ Qual Saf*. Oct 18, 2024;33(11):748-754. [doi: [10.1136/bmjqs-2023-016690](https://doi.org/10.1136/bmjqs-2023-016690)] [Medline: [39048292](https://pubmed.ncbi.nlm.nih.gov/39048292/)]
20. Wang Z, Wang H, Danek B, et al. A perspective for adapting generalist AI to specialized medical AI applications and their challenges. *NPJ Digit Med*. 2025;8(1):429. [doi: [10.1038/s41746-025-01789-7](https://doi.org/10.1038/s41746-025-01789-7)]
21. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. Jan 2025;31(1):60-69. [doi: [10.1038/s41591-024-03425-5](https://doi.org/10.1038/s41591-024-03425-5)] [Medline: [39779929](https://pubmed.ncbi.nlm.nih.gov/39779929/)]
22. Miao BY, Chen IY, Williams CYK, et al. The MI-CLAIM-GEN checklist for generative artificial intelligence in health. *Nat Med*. May 2025;31(5):1394-1398. [doi: [10.1038/s41591-024-03470-0](https://doi.org/10.1038/s41591-024-03470-0)] [Medline: [39915678](https://pubmed.ncbi.nlm.nih.gov/39915678/)]
23. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med*. Dec 2023;29(12):2988. [doi: [10.1038/s41591-023-02656-2](https://doi.org/10.1038/s41591-023-02656-2)] [Medline: [37957381](https://pubmed.ncbi.nlm.nih.gov/37957381/)]

24. Tripathi S, Alkhulaifat D, Doo FX, et al. Development, evaluation, and assessment of large language models (DEAL) checklist: a technical report. *NEJM AI*. May 22, 2025;2(6):AIP2401106. [doi: [10.1056/AIP2401106](https://doi.org/10.1056/AIP2401106)]
25. Chandler J, Cumpston M, Li T, Page MJ. *Cochrane Handbook for Systematic Reviews of Interventions* Version. Vol 6. Cochrane; :5.
26. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. Mar 29, 2021;372:n71. [doi: [10.1136/bmj.n71](https://doi.org/10.1136/bmj.n71)] [Medline: [33782057](https://pubmed.ncbi.nlm.nih.gov/33782057/)]
27. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ*. Jan 16, 2020;368:l6890. [doi: [10.1136/bmj.l6890](https://doi.org/10.1136/bmj.l6890)] [Medline: [31948937](https://pubmed.ncbi.nlm.nih.gov/31948937/)]
28. Teo ZL, Thirunavukarasu AJ, Elangovan K, et al. Generative artificial intelligence in medicine. *Nat Med*. Oct 2025;31(10):3270-3282. [doi: [10.1038/s41591-025-03983-2](https://doi.org/10.1038/s41591-025-03983-2)] [Medline: [41053447](https://pubmed.ncbi.nlm.nih.gov/41053447/)]
29. He R, Cao J, Tan T. Generative artificial intelligence: a historical perspective. *Natl Sci Rev*. May 2025;12(5):nwaf050. [doi: [10.1093/nsr/nwaf050](https://doi.org/10.1093/nsr/nwaf050)] [Medline: [40191253](https://pubmed.ncbi.nlm.nih.gov/40191253/)]
30. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. *Bus Inf Syst Eng*. Feb 2024;66(1):111-126. [doi: [10.1007/s12599-023-00834-7](https://doi.org/10.1007/s12599-023-00834-7)]
31. Ronge R, Maier M, Rathgeber B. Towards a definition of generative artificial intelligence. *Philos Technol*. Mar 2025;38(1). [doi: [10.1007/s13347-025-00863-y](https://doi.org/10.1007/s13347-025-00863-y)]
32. Haddaway NR, Grainger MJ, Gray CT. Citationchaser: an r package for forward and backward citations chasing in academic searching. Zenodo; 2021. [doi: [10.5281/zenodo.4543513](https://doi.org/10.5281/zenodo.4543513)]
33. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Syst Rev*. Jan 26, 2021;10(1):39. [doi: [10.1186/s13643-020-01542-z](https://doi.org/10.1186/s13643-020-01542-z)] [Medline: [33499930](https://pubmed.ncbi.nlm.nih.gov/33499930/)]
34. Swamy P. Personalized nutrition prescription generated by AI platform improves cardiometabolic health. *Int J Life Sci Biotechnol Pharma Res*. 2025;14(7):1554-1558. [doi: [10.69605/ijlbpr_14.7.2025.259](https://doi.org/10.69605/ijlbpr_14.7.2025.259)]
35. Cheng ST, Ng PHF. Intervention without borders - an automated self-guided AI-enhanced psychoeducation intervention for dementia caregivers: parallel-group randomized waitlist-controlled trial. *Am J Geriatr Psychiatry*. Oct 2026;34(10):1284-1296. [doi: [10.1016/j.jagp.2026.01.001](https://doi.org/10.1016/j.jagp.2026.01.001)] [Medline: [41644333](https://pubmed.ncbi.nlm.nih.gov/41644333/)]
36. Cheng ST, Ng PHF. Multiple therapeutic mechanisms of an automated self-guided AI-enhanced psychoeducation intervention for dementia family caregivers: randomized waitlist-controlled trial. *Am J Geriatr Psychiatry*. Nov 2026;34(11):1423-1434. [doi: [10.1016/j.jagp.2026.02.003](https://doi.org/10.1016/j.jagp.2026.02.003)] [Medline: [41925437](https://pubmed.ncbi.nlm.nih.gov/41925437/)]
37. Zhao H, Bai Y, Chen L, et al. Generative AI-based low-dose digital subtraction angiography for intra-operative radiation dose reduction: a randomized controlled trial. *Nat Med*. Jan 2026;32(1):288-296. [doi: [10.1038/s41591-025-04042-6](https://doi.org/10.1038/s41591-025-04042-6)]
38. Cao Y, Lu W, Shi R, et al. Performance of popular large language models in glaucoma patient education: a randomized controlled study. *Adv Ophthalmol Pract Res*. 2025;5(2):88-94. [doi: [10.1016/j.aopr.2024.12.002](https://doi.org/10.1016/j.aopr.2024.12.002)] [Medline: [40162329](https://pubmed.ncbi.nlm.nih.gov/40162329/)]
39. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open*. Oct 1, 2024;7(10):e2440969. [doi: [10.1001/jamanetworkopen.2024.40969](https://doi.org/10.1001/jamanetworkopen.2024.40969)] [Medline: [39466245](https://pubmed.ncbi.nlm.nih.gov/39466245/)]
40. Wan P, Huang Z, Tang W, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat Med*. Oct 2024;30(10):2878-2885. [doi: [10.1038/s41591-024-03148-7](https://doi.org/10.1038/s41591-024-03148-7)] [Medline: [39009780](https://pubmed.ncbi.nlm.nih.gov/39009780/)]
41. Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's ability to assist with clinical documentation: a randomized controlled trial. *J Am Acad Orthop Surg*. Feb 1, 2024;32(3):123-129. [doi: [10.5435/JAAOS-D-23-00474](https://doi.org/10.5435/JAAOS-D-23-00474)] [Medline: [37976385](https://pubmed.ncbi.nlm.nih.gov/37976385/)]
42. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. *Nature*. Jun 2025;642(8067):451-457. [doi: [10.1038/s41586-025-08869-4](https://doi.org/10.1038/s41586-025-08869-4)] [Medline: [40205049](https://pubmed.ncbi.nlm.nih.gov/40205049/)]
43. Gorenstein A, Weisblat Y, Khateb M, et al. AI-based EMG reporting: a randomized controlled trial. *J Neurol*. Aug 22, 2025;272(9):586. [doi: [10.1007/s00415-025-13261-3](https://doi.org/10.1007/s00415-025-13261-3)] [Medline: [40844612](https://pubmed.ncbi.nlm.nih.gov/40844612/)]
44. Wu Y, Chen X, Zhang W, et al. ChatMyopia: an AI agent for myopia-related consultation in primary eye care settings. *iScience*. Nov 2025;28(11):113768. [doi: [10.1016/j.isci.2025.113768](https://doi.org/10.1016/j.isci.2025.113768)]
45. Ke YH, Yang Ong BS, Jin L, et al. Clinical and economic impact of a large language model in perioperative medicine: a randomized crossover trial. *NPJ Digit Med*. Jul 21, 2025;8(1):462. [doi: [10.1038/s41746-025-01858-x](https://doi.org/10.1038/s41746-025-01858-x)] [Medline: [40691284](https://pubmed.ncbi.nlm.nih.gov/40691284/)]
46. Tailor PD, D'Souza HS, Castillejo Becerra CM, et al. Evaluation of AI summaries on interdisciplinary understanding of ophthalmology notes. *JAMA Ophthalmol*. May 1, 2025;143(5):410-419. [doi: [10.1001/jamaophthalmol.2025.0351](https://doi.org/10.1001/jamaophthalmol.2025.0351)] [Medline: [40178837](https://pubmed.ncbi.nlm.nih.gov/40178837/)]

47. Ye X, Shan X, Tu Y, Zhang Y. Examining the efficacy of large language models for mitigating depression and anxiety among Chinese students: a randomized controlled trial. *Comput Inform Nurs*. Sep 1, 2025;43(9):e01349. [doi: [10.1097/CIN.0000000000001349](https://doi.org/10.1097/CIN.0000000000001349)] [Medline: [40704906](https://pubmed.ncbi.nlm.nih.gov/40704906/)]
48. Sabour S, Zhang W, Xiao X, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. *Front Digit Health*. 2023;5:1133987. [doi: [10.3389/fdgth.2023.1133987](https://doi.org/10.3389/fdgth.2023.1133987)] [Medline: [37214342](https://pubmed.ncbi.nlm.nih.gov/37214342/)]
49. Aydin F, Yildirim ÖT, Aydin AH, Murat B, Basaran CH. Comparison of artificial intelligence-assisted informed consent obtained before coronary angiography with the conventional method: Medical competence and ethical assessment. *Digit HEALTH*. 2023;9:20552076231218141. [doi: [10.1177/20552076231218141](https://doi.org/10.1177/20552076231218141)] [Medline: [38047164](https://pubmed.ncbi.nlm.nih.gov/38047164/)]
50. Tai-Seale M, Baxter SL, Vaida F, et al. AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw Open*. Apr 1, 2024;7(4):e246565. [doi: [10.1001/jamanetworkopen.2024.6565](https://doi.org/10.1001/jamanetworkopen.2024.6565)] [Medline: [38619840](https://pubmed.ncbi.nlm.nih.gov/38619840/)]
51. Yahagi M, Hiruta R, Miyauchi C, Tanaka S, Taguchi A, Yaguchi Y. Comparison of Conventional Anesthesia Nurse Education and an Artificial Intelligence Chatbot (ChatGPT) intervention on preoperative anxiety: A randomized controlled trial. *J Perianesth Nurs*. Oct 2024;39(5):767-771. [doi: [10.1016/j.jopan.2023.12.005](https://doi.org/10.1016/j.jopan.2023.12.005)] [Medline: [38520470](https://pubmed.ncbi.nlm.nih.gov/38520470/)]
52. Chung D, Sidhom K, Dhillon H, et al. Real-world utility of ChatGPT in pre-vasectomy counselling, a safe and efficient practice: a prospective single-centre clinical study. *World J Urol*. Dec 14, 2024;43(1):32. [doi: [10.1007/s00345-024-05385-4](https://doi.org/10.1007/s00345-024-05385-4)] [Medline: [39673635](https://pubmed.ncbi.nlm.nih.gov/39673635/)]
53. Ebnali Harari R, Altaweel A, Ahram T, Keehner M, Shokoohi H. A randomized controlled trial on evaluating clinician-supervised generative AI for decision support. *Int J Med Inform*. Mar 2025;195:105701. [doi: [10.1016/j.ijmedinf.2024.105701](https://doi.org/10.1016/j.ijmedinf.2024.105701)] [Medline: [39631268](https://pubmed.ncbi.nlm.nih.gov/39631268/)]
54. Hou Z, Wu Z, Qu Z, et al. A vaccine chatbot intervention for parents to improve HPV vaccination uptake among middle school girls: A cluster randomized trial. *Nat Med*. Jun 2025;31(6):1855-1862. [doi: [10.1038/s41591-025-03618-6](https://doi.org/10.1038/s41591-025-03618-6)] [Medline: [40195450](https://pubmed.ncbi.nlm.nih.gov/40195450/)]
55. Lukac PJ, Turner W, Vangala S, et al. Ambient AI scribes in clinical practice: A randomized trial. *NEJM AI*. Dec 2025;2(12). [doi: [10.1056/aioa2501000](https://doi.org/10.1056/aioa2501000)] [Medline: [41497288](https://pubmed.ncbi.nlm.nih.gov/41497288/)]
56. Wu Y, Qian B, Li T, et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. *Nat Med*. Oct 2025;31(10):3404-3413. [doi: [10.1038/s41591-025-03900-7](https://doi.org/10.1038/s41591-025-03900-7)] [Medline: [40877476](https://pubmed.ncbi.nlm.nih.gov/40877476/)]
57. Berzolla E, Gosnell GG, Chen L, Vonck C, Alaia E, Meislin R. Artificial intelligence large language models improve patient comprehension of radiologist magnetic resonance imaging reports. *Arthroscopy*. Nov 2025;41(11):4607-4614. [doi: [10.1016/j.arthro.2025.04.033](https://doi.org/10.1016/j.arthro.2025.04.033)] [Medline: [40288466](https://pubmed.ncbi.nlm.nih.gov/40288466/)]
58. Kerimoglu Yildiz G, Turk Delibalta R, Coktay Z. Artificial intelligence-assisted chatbot: impact on breastfeeding outcomes and maternal anxiety. *BMC Pregnancy Childbirth*. May 30, 2025;25(1):631. [doi: [10.1186/s12884-025-07753-3](https://doi.org/10.1186/s12884-025-07753-3)] [Medline: [40448061](https://pubmed.ncbi.nlm.nih.gov/40448061/)]
59. Aydemir U. ChatGPT-delivered physical activity intervention for children with autism spectrum disorder: Pre-post feasibility study. *JMIR Hum Factors*. Jun 11, 2025;12:e71119. [doi: [10.2196/71119](https://doi.org/10.2196/71119)] [Medline: [40498657](https://pubmed.ncbi.nlm.nih.gov/40498657/)]
60. Gan W, Ouyang J, She G, et al. ChatGPT's role in alleviating anxiety in total knee arthroplasty consent process: a randomized controlled trial pilot study. *Int J Surg*. Mar 1, 2025;111(3):2546-2557. [doi: [10.1097/JS9.0000000000002223](https://doi.org/10.1097/JS9.0000000000002223)] [Medline: [39903546](https://pubmed.ncbi.nlm.nih.gov/39903546/)]
61. Chen C, Lam KT, Yip KM, et al. Comparison of an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: Pilot randomized controlled trial. *JMIR Hum Factors*. Mar 6, 2025;12:e65785. [doi: [10.2196/65785](https://doi.org/10.2196/65785)] [Medline: [40048637](https://pubmed.ncbi.nlm.nih.gov/40048637/)]
62. Weuthen FA, Otte N, Krabbe H, Kraus T, Krabbe J. Comparison of ChatGPT and internet research for clinical research and decision-making in occupational medicine: Randomized controlled trial. *JMIR Form Res*. May 20, 2025;9:e63857. [doi: [10.2196/63857](https://doi.org/10.2196/63857)] [Medline: [40393042](https://pubmed.ncbi.nlm.nih.gov/40393042/)]
63. Zhang H, Wang X, Luo H, et al. Comparison of preoperative education by artificial intelligence versus traditional physicians in perioperative management of urolithiasis surgery: a prospective single-blind randomized controlled trial conducted in China. *Front Med*. ;12. [doi: [10.3389/fmed.2025.1543630](https://doi.org/10.3389/fmed.2025.1543630)]
64. Akdogan O, Uyar GC, Yesilbas E, et al. Effect of a ChatGPT-based digital counseling intervention on anxiety and depression in patients with cancer: A prospective, randomized trial. *Eur J Cancer*. May 15, 2025;221:115408. [doi: [10.1016/j.ejca.2025.115408](https://doi.org/10.1016/j.ejca.2025.115408)] [Medline: [40215593](https://pubmed.ncbi.nlm.nih.gov/40215593/)]
65. Zhao Y, Qian W, Chen Y, et al. Effect of an AI agent trained on a large language model (LLM) as an intervention for depression and anxiety symptoms in young adults: A 28-day randomized controlled trial. *Appl Psychol Health Well Being*. Oct 2025;17(5):e70067. [doi: [10.1111/aphw.70067](https://doi.org/10.1111/aphw.70067)] [Medline: [40910958](https://pubmed.ncbi.nlm.nih.gov/40910958/)]
66. Lai HL, Chen CI, Liaw GH, Lu LC, Huang CY. Effectiveness of a chatbot-based internet-support-coping-caring intervention in improving sleep quality and work-related outcomes among nurses in Taiwan during the COVID-19

- pandemic: A pilot study. *J Nurs Res*. Oct 1, 2025;33(5):e411. [doi: [10.1097/jnr.0000000000000691](https://doi.org/10.1097/jnr.0000000000000691)] [Medline: [40748237](https://pubmed.ncbi.nlm.nih.gov/40748237/)]
67. Anand E, Ghersin I, Lingam G, et al. Enhancing patient understanding of perianal fistula MRI findings using ChatGPT: A randomized, single centre study. *Diagnostics (Basel)*. Dec 25, 2025;16(1):72. [doi: [10.3390/diagnostics16010072](https://doi.org/10.3390/diagnostics16010072)] [Medline: [41515567](https://pubmed.ncbi.nlm.nih.gov/41515567/)]
68. Islam S. Evaluating the impact of AI-generated educational content on patient understanding and anxiety in endodontics and restorative dentistry: A comparative study. *BMC Oral Health*. May 7, 2025;25(1):689. [doi: [10.1186/s12903-025-06069-0](https://doi.org/10.1186/s12903-025-06069-0)] [Medline: [40335999](https://pubmed.ncbi.nlm.nih.gov/40335999/)]
69. Busch F, Kaibel L, Nguyen H, et al. Evaluation of a retrieval-augmented generation-powered chatbot for Pre-CT informed consent: A prospective comparative study. *J Imaging Inform Med*. Dec 2025;38(6):4312-4323. [doi: [10.1007/s10278-025-01483-w](https://doi.org/10.1007/s10278-025-01483-w)] [Medline: [40119020](https://pubmed.ncbi.nlm.nih.gov/40119020/)]
70. Donner S, Knauer P, Kienzle A, et al. Exploring patient perspectives, engagement, and output quality in doctor-supervised use of artificial intelligence during informed consent consultation with ChatGPT and retrieval augmented generation (RAG): Quantitative exploratory study. *J Med Internet Res*. Oct 22, 2025;27:e73717. [doi: [10.2196/73717](https://doi.org/10.2196/73717)] [Medline: [41124695](https://pubmed.ncbi.nlm.nih.gov/41124695/)]
71. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. *Nat Med*. Apr 2025;31(4):1233-1238. [doi: [10.1038/s41591-024-03456-y](https://doi.org/10.1038/s41591-024-03456-y)] [Medline: [39910272](https://pubmed.ncbi.nlm.nih.gov/39910272/)]
72. Stephan D, Bertsch AS, Schumacher S, et al. Improving patient communication by simplifying AI-generated dental radiology reports with ChatGPT: Comparative study. *J Med Internet Res*. Jun 9, 2025;27:e73337. [doi: [10.2196/73337](https://doi.org/10.2196/73337)] [Medline: [40489773](https://pubmed.ncbi.nlm.nih.gov/40489773/)]
73. Allen JW, Rodríguez Hannikainen I, Savulescu J, Wilkinson D, Earp BD. Is Consent-GPT valid? Public attitudes to generative AI use in surgical consent. *AI Soc*. Mar 2026;41(3):2637-2655. [doi: [10.1007/s00146-025-02644-9](https://doi.org/10.1007/s00146-025-02644-9)] [Medline: [41189744](https://pubmed.ncbi.nlm.nih.gov/41189744/)]
74. Samardzic D, Curkovic J, Okmazic D, et al. Patients prefer human empathy, but not always human wording: A single-blind within-subject trial of GPT-generated vs. clinician discharge texts in emergency ophthalmology. *Clin Pract*. Nov 14, 2025;15(11):208. [doi: [10.3390/clinpract15110208](https://doi.org/10.3390/clinpract15110208)] [Medline: [41294639](https://pubmed.ncbi.nlm.nih.gov/41294639/)]
75. Goh E, Bunning B, Khoong EC, et al. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. *Commun Med (Lond)*. Mar 4, 2025;5(1):59. [doi: [10.1038/s43856-025-00781-2](https://doi.org/10.1038/s43856-025-00781-2)] [Medline: [40038550](https://pubmed.ncbi.nlm.nih.gov/40038550/)]
76. Gupta A, Singh S, Malhotra H, et al. Provision of radiology reports simplified with large language models to patients with cancer: Impact on patient satisfaction. *JCO Clin Cancer Inform*. Jan 2025;9(9):e2400166. [doi: [10.1200/CCI-24-00166](https://doi.org/10.1200/CCI-24-00166)] [Medline: [39879570](https://pubmed.ncbi.nlm.nih.gov/39879570/)]
77. Indrayanti I, Salsabila AK, Amrita V, Alhaddad MM, Saskia AB, Ramadhani DP. PsyBot: A randomized controlled trial of WhatsApp-based psychological first aid to reduce loneliness among 18–22-year-old students in Yogyakarta, Indonesia. *SSM - Mental Health*. Dec 2025;8:100504. [doi: [10.1016/j.ssmmh.2025.100504](https://doi.org/10.1016/j.ssmmh.2025.100504)]
78. Campellone TR, Flom M, Montgomery RM, et al. Safety and user experience of a generative artificial intelligence digital mental health intervention: Exploratory randomized controlled trial. *J Med Internet Res*. May 23, 2025;27:e67365. [doi: [10.2196/67365](https://doi.org/10.2196/67365)] [Medline: [40408143](https://pubmed.ncbi.nlm.nih.gov/40408143/)]
79. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. *Nature*. Jun 2025;642(8067):442-450. [doi: [10.1038/s41586-025-08866-7](https://doi.org/10.1038/s41586-025-08866-7)] [Medline: [40205050](https://pubmed.ncbi.nlm.nih.gov/40205050/)]
80. Chung S, Giuffrè M, Rajashekar N, et al. Usability and adoption in a randomized trial of GutGPT a GenAI tool for gastrointestinal bleeding. *NPJ Digit Med*. Aug 18, 2025;8(1):527. [doi: [10.1038/s41746-025-01896-5](https://doi.org/10.1038/s41746-025-01896-5)] [Medline: [40825997](https://pubmed.ncbi.nlm.nih.gov/40825997/)]
81. Koegel LK, Ponder E, Bruzzese T, et al. Using artificial intelligence to improve empathetic statements in autistic adolescents and adults: A randomized clinical trial. *J Autism Dev Disord*. Jul 2026;56(7):2513-2529. [doi: [10.1007/s10803-025-06734-x](https://doi.org/10.1007/s10803-025-06734-x)] [Medline: [39954218](https://pubmed.ncbi.nlm.nih.gov/39954218/)]
82. Rollwage M, McFadyen J, Juchems K, et al. A cognitive layer architecture to support large-language model performance in psychotherapy interactions. *Nat Med*. May 2026;32(5):1717-1725. [doi: [10.1038/s41591-026-04278-w](https://doi.org/10.1038/s41591-026-04278-w)] [Medline: [41820675](https://pubmed.ncbi.nlm.nih.gov/41820675/)]
83. Li S, Li Y, Zhou S, et al. A community-codesigned LLM-powered chatbot for primary care: A randomized controlled trial. *Nat Health*. 2026;1(2):238-250. [doi: [10.1038/s44360-025-00021-w](https://doi.org/10.1038/s44360-025-00021-w)] [Medline: [41659358](https://pubmed.ncbi.nlm.nih.gov/41659358/)]
84. Liu Z, Orouji F, Yang F, et al. A fine-tuned large language model chatbot for multi-scenario radiology cancer care: Randomized controlled trial on interaction optimization, emotional support, and provider burnout reduction. *J Transl Med*. ;24(1). [doi: [10.1186/s12967-026-07738-6](https://doi.org/10.1186/s12967-026-07738-6)]

85. O'Sullivan JW, Palepu A, Saab K, et al. A large language model for complex cardiology care. *Nat Med*. Feb 2026;32(2):616-623. [doi: [10.1038/s41591-025-04190-9](https://doi.org/10.1038/s41591-025-04190-9)] [Medline: [41652123](https://pubmed.ncbi.nlm.nih.gov/41652123/)]
86. Li J, Zhang Y, Zhang Z, et al. A randomized controlled trial of a WeChat-based artificial intelligence agent for postoperative care in orthopedic patients. *npj Digit Med*. ;9(1). [doi: [10.1038/s41746-025-02269-8](https://doi.org/10.1038/s41746-025-02269-8)]
87. Yang X, Huo Y, Wang Y, et al. AI-enhanced versus clinician-guided expressive writing for improving psychological well-being in cancer survivors: A randomized controlled trial. *Psychooncology*. Jan 2026;35(1):e70386. [doi: [10.1002/pon.70386](https://doi.org/10.1002/pon.70386)] [Medline: [41539978](https://pubmed.ncbi.nlm.nih.gov/41539978/)]
88. Tao X, Zhou S, Ding K, et al. An LLM chatbot to facilitate primary-to-specialist care transitions: A randomized controlled trial. *Nat Med*. Mar 2026;32(3):934-942. [doi: [10.1038/s41591-025-04176-7](https://doi.org/10.1038/s41591-025-04176-7)] [Medline: [41555035](https://pubmed.ncbi.nlm.nih.gov/41555035/)]
89. Koç U, Güneş YC, Çolakoglu MN, et al. ChatGPT-generated informed consent forms in interventional radiology: A randomized controlled evaluation of comprehension, attitudinal responses, and readability. *Eur J Radiol*. Feb 2026;195:112624. [doi: [10.1016/j.ejrad.2025.112624](https://doi.org/10.1016/j.ejrad.2025.112624)] [Medline: [41483678](https://pubmed.ncbi.nlm.nih.gov/41483678/)]
90. Navarese EP, Leader JH, Markides RIL, et al. Design and architecture of a generative-AI-supported, nonphysician-delivered model for GDMT optimization in HFrEF. *JACC Adv*. Mar 2026;5(3):102588. [doi: [10.1016/j.jacadv.2026.102588](https://doi.org/10.1016/j.jacadv.2026.102588)] [Medline: [41670585](https://pubmed.ncbi.nlm.nih.gov/41670585/)]
91. Nguyen QC, Norell EM, Mane H, et al. Developing an artificial intelligence-powered question-and-answer chatbot with English-Spanish capabilities for new mothers. *JAMIA Open*. Feb 2026;9(1):ooaf169. [doi: [10.1093/jamiaopen/ooaf169](https://doi.org/10.1093/jamiaopen/ooaf169)] [Medline: [41502532](https://pubmed.ncbi.nlm.nih.gov/41502532/)]
92. Sun Y, Xu S, Jin H, et al. Effectiveness of AI-assisted patient health education using voice cloning and ChatGPT: Prospective randomized controlled trial. *J Med Internet Res*. ;28:e81387. [doi: [10.2196/81387](https://doi.org/10.2196/81387)]
93. Ouyang H, Lin G, Li Y, et al. Evaluating and validating large language models for health education on developmental dysplasia of the hip: 2-Phase study with expert ratings and a pilot randomized controlled trial. *J Med Internet Res*. Jan 19, 2026;28:e73326. [doi: [10.2196/73326](https://doi.org/10.2196/73326)] [Medline: [41554120](https://pubmed.ncbi.nlm.nih.gov/41554120/)]
94. Everett SS, Bunning BJ, Jain P, et al. From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. *NPJ Digit Med*. Mar 18, 2026;9(1):409. [doi: [10.1038/s41746-026-02545-1](https://doi.org/10.1038/s41746-026-02545-1)] [Medline: [41851268](https://pubmed.ncbi.nlm.nih.gov/41851268/)]
95. Rahrish A, Braun J, Ganter MT, et al. Generative AI in preanesthetic consultations: Effects on efficiency, documentation workload, quality, and physician-patient interaction: A simulation trial. *J Clin Anesth*. Mar 2026;110:112163. [doi: [10.1016/j.jclinane.2026.112163](https://doi.org/10.1016/j.jclinane.2026.112163)] [Medline: [41719826](https://pubmed.ncbi.nlm.nih.gov/41719826/)]
96. Yıldız HA. Human doctors or AI? Evaluating patient satisfaction in urinary stone disease consultations. *J Surg Res*. Jan 2026;317:181-188. [doi: [10.1016/j.jss.2025.11.026](https://doi.org/10.1016/j.jss.2025.11.026)] [Medline: [41385971](https://pubmed.ncbi.nlm.nih.gov/41385971/)]
97. Camathias C, Papp K, Betschart P, Speth B, Valderrabano V, Ammann E. Human medical documentation significantly outperforms ChatGPT-4o in critical clinical dimensions: A blinded comparative assessment in paediatric orthopaedics. *Knee Surg Sports Traumatol Arthrosc*. May 2026;34(5):1910-1916. [doi: [10.1002/ksa.70357](https://doi.org/10.1002/ksa.70357)] [Medline: [41762622](https://pubmed.ncbi.nlm.nih.gov/41762622/)]
98. Holderried F, Sonanini A, Stegemann-Philippis C, et al. Impact of GPT-4-generated discharge letters on patients' medical comprehension: Prospective crossover study. *J Med Internet Res*. Feb 26, 2026;28:e81243. [doi: [10.2196/81243](https://doi.org/10.2196/81243)] [Medline: [41746691](https://pubmed.ncbi.nlm.nih.gov/41746691/)]
99. Chen AH, Rudin RS, Levine DM, Mehrotra A. Improving patient understanding of radiology reports using generative artificial intelligence: A vignette study of 2000 US adults. *J Am Med Inform Assoc*. Feb 1, 2026;33(2):326-335. [doi: [10.1093/jamia/ocaf187](https://doi.org/10.1093/jamia/ocaf187)] [Medline: [41241949](https://pubmed.ncbi.nlm.nih.gov/41241949/)]
100. McFadyen J, Habicht J, Dina LM, Harper R, Hauser TU, Rollwage M. Increasing engagement with cognitive-behavioral therapy (CBT) using generative AI: A randomized controlled trial (RCT). *Commun Med (Lond)*. Jan 15, 2026;6(1):129. [doi: [10.1038/s43856-025-01321-8](https://doi.org/10.1038/s43856-025-01321-8)] [Medline: [41540194](https://pubmed.ncbi.nlm.nih.gov/41540194/)]
101. Bean AM, Payne RE, Parsons G, et al. Reliability of LLMs as medical assistants for the general public: A randomized preregistered study. *Nat Med*. Feb 2026;32(2):609-615. [doi: [10.1038/s41591-025-04074-y](https://doi.org/10.1038/s41591-025-04074-y)] [Medline: [41663592](https://pubmed.ncbi.nlm.nih.gov/41663592/)]
102. Mutlu Sütcüoğlu B, Öztöpm MM, Hansu K. The impact of ChatGPT on preoperative anxiety and postoperative depression in cesarean section: A prospective randomized trial. *BMC Pregnancy Childbirth*. Apr 1, 2026;26(1):517. [doi: [10.1186/s12884-026-09011-6](https://doi.org/10.1186/s12884-026-09011-6)] [Medline: [41923223](https://pubmed.ncbi.nlm.nih.gov/41923223/)]
103. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*. Mar 27, 2025;2(4). [doi: [10.1056/AIoa2400802](https://doi.org/10.1056/AIoa2400802)]
104. Song Y, Yan T, Jia F, Chen L, Li H. Developing generative AI for value co-creation: An intervention-based randomized field experiment in a healthcare context. *J of Ops Management*. Mar 2026;72(2):197-232. [doi: [10.1002/joom.70012](https://doi.org/10.1002/joom.70012)]
105. Qazi IA, Ali A, Khawaja AU, Akhtar MJ, Sheikh AZ, Alizai MH. Large language model diagnostic assistance for physicians in a lower-middle-income country: A randomized controlled trial. *Nat Health*. ;1(2):198-205. [doi: [10.1038/s44360-025-00007-8](https://doi.org/10.1038/s44360-025-00007-8)]

106. Etagiuri A, Chaudhry D, van der Merwe J, Nickol M, van der Merwe JM. Evaluating the impact of artificial intelligence scribes on clinical workflow and documentation quality: A randomized controlled trial. *J Arthroplasty*. Apr 2, 2026;S0883-5403(26)00272-X. [doi: [10.1016/j.arth.2026.03.069](https://doi.org/10.1016/j.arth.2026.03.069)] [Medline: [41935758](https://pubmed.ncbi.nlm.nih.gov/41935758/)]
107. Shin HS, Williams H, Braykov N, Jahan A, Meller J, Orenstein EW. The influence of artificial intelligence scribes on clinician experience and efficiency among pediatric subspecialists: A rapid, randomized quality improvement trial. *Appl Clin Inform*. Aug 2025;16(4):1041-1052. [doi: [10.1055/a-2657-8087](https://doi.org/10.1055/a-2657-8087)] [Medline: [40675605](https://pubmed.ncbi.nlm.nih.gov/40675605/)]
108. Kakaday R, Herrera EZ, Coskey O, Hertel AW, Kaiser P. The STREAMLINE pilot study on time reduction and efficiency in AI-mediated logging for improved note-taking experience. *Appl Clin Inform*. May 2025;16(3):614-621. [doi: [10.1055/a-2559-5791](https://doi.org/10.1055/a-2559-5791)] [Medline: [40097146](https://pubmed.ncbi.nlm.nih.gov/40097146/)]
109. Schick T, Dwivedi-Yu J, Dessi R, et al. Toolformer: language models can teach themselves to use tools. Presented at: Advances in Neural Information Processing Systems 36; Dec 10-16, 2023:68539-68551; New Orleans, Louisiana. 2023.URL: <http://www.proceedings.com/75280.html> [Accessed 2026-10-12] [doi: [10.52202/075280-2997](https://doi.org/10.52202/075280-2997)]
110. Rebitschek FG, Carella A, Kohlrausch-Pazin S, Zitzmann M, Steckelberg A, Wilhelm C. Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information. *NPJ Digit Med*. Jun 9, 2025;8(1):343. [doi: [10.1038/s41746-025-01752-6](https://doi.org/10.1038/s41746-025-01752-6)] [Medline: [40490558](https://pubmed.ncbi.nlm.nih.gov/40490558/)]
111. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature*. Oct 2020;586(7829):E14-E16. [doi: [10.1038/s41586-020-2766-y](https://doi.org/10.1038/s41586-020-2766-y)] [Medline: [33057217](https://pubmed.ncbi.nlm.nih.gov/33057217/)]
112. Deprecations. OpenAI. 2026. URL: <https://developers.openai.com/api/docs/deprecations> [Accessed 2026-10-12]
113. Yang C, Wang X, Lu Y, Liu H, Le QV, Zhou D, et al. Large language models as optimizers. *ArXiv*. Preprint posted online on Apr 15, 2024. [doi: [10.48550/arXiv.2309.03409](https://doi.org/10.48550/arXiv.2309.03409)]
114. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. Sep 2024;30(9):2613-2622. [doi: [10.1038/s41591-024-03097-1](https://doi.org/10.1038/s41591-024-03097-1)] [Medline: [38965432](https://pubmed.ncbi.nlm.nih.gov/38965432/)]
115. Korn A, Zaruchas L, Arora C, et al. Reporting LLM prompting in automated software engineering: a guideline based on current practices and expectations. Presented at: FORGE '26; Apr 12-13, 2026:184-194; Rio de Janeiro, Brazil. Apr 12, 2026.URL: <https://dl.acm.org/doi/proceedings/10.1145/3793655> [Accessed 2026-10-12] [doi: [10.1145/3793655.3793719](https://doi.org/10.1145/3793655.3793719)]
116. Warren CJ, Edmonds VS, Payne NG, et al. Prompt matters: evaluation of large language model chatbot responses related to Peyronie's disease. *Sex Med*. Aug 2024;12(4):qfae055. [doi: [10.1093/sexmed/qfae055](https://doi.org/10.1093/sexmed/qfae055)] [Medline: [39257694](https://pubmed.ncbi.nlm.nih.gov/39257694/)]
117. Salemi A, Zamani H. Evaluating retrieval quality in retrieval-augmented generation. Presented at: SIGIR 2024; Jul 14-18, 2024; Washington DC USA. Jul 10, 2024.URL: <https://dl.acm.org/doi/proceedings/10.1145/3626772> [Accessed 2026-10-12] [doi: [10.1145/3626772.3657957](https://doi.org/10.1145/3626772.3657957)]
118. Ma X, Gong Y, He P, Zhao H, Duan N. Query rewriting in retrieval-augmented large language models. Oct 23, 2023. [doi: [10.48550/arXiv.2305.14283](https://doi.org/10.48550/arXiv.2305.14283)]
119. Health insurance portability and accountability act of 1996. Office of the Assistant Secretary for Planning and Evaluation (ASPE). 2024. URL: <https://aspe.hhs.gov/reports/health-insurance-portability-accountability-act-1996> [Accessed 2026-10-25]
120. Li J. Security Implications of AI Chatbots in Health Care. *J Med Internet Res*. 2023;25:e47551. [doi: [10.2196/47551](https://doi.org/10.2196/47551)]
121. Data, privacy, and security for models sold by azure in microsoft foundry. Microsoft Learn. 2026. URL: <https://learn.microsoft.com/en-us/azure/foundry/responsible-ai/openai/data-privacy> [Accessed 2026-10-12]
122. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation) (text with EEA relevance). EUR-Lex. 2016. URL: <https://eur-lex.europa.eu/eli/reg/2016/679/oj> [Accessed 2026-10-25]
123. ChatGPT-release notes. OpenAI. URL: <https://help.openai.com/en/articles/6825453-chatgpt-release-notes> [Accessed 2026-10-12]
124. Introducing meta llama 3: the most capable openly available LLM to date. Meta. URL: <https://ai.meta.com/blog/meta-llama-3> [Accessed 2025-10-02]
125. Chen L, Zaharia M, Zou J. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*. 2024;6(2). [doi: [10.1162/99608f92.5317da47](https://doi.org/10.1162/99608f92.5317da47)]
126. Nakano R, Hilton J, Balaji S, Wu J, Ouyang L, Kim C, et al. Webgpt: browser-assisted question-answering with human feedback. *arXiv*. Preprint posted online on Jun 1, 2022. [doi: [10.48550/arXiv.2112.09332](https://doi.org/10.48550/arXiv.2112.09332)]
127. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. Oct 10, 2023;3(1):141. [doi: [10.1038/s43856-023-00370-1](https://doi.org/10.1038/s43856-023-00370-1)] [Medline: [37816837](https://pubmed.ncbi.nlm.nih.gov/37816837/)]
128. Chuang YN, Tang R, Jiang X, Hu X. SPeC: a soft prompt-based calibration on performance variability of large language model in clinical notes summarization. *J Biomed Inform*. Mar 2024;151:104606. [doi: [10.1016/j.jbi.2024.104606](https://doi.org/10.1016/j.jbi.2024.104606)] [Medline: [38325698](https://pubmed.ncbi.nlm.nih.gov/38325698/)]
129. Jarrett PC, Hill J, Howell M, et al. Piloting temperature-driven variability in emergency diagnostic accuracy using a leading large language model. *Cureus*. Oct 2025;17(10):e94476. [doi: [10.7759/cureus.94476](https://doi.org/10.7759/cureus.94476)] [Medline: [41235003](https://pubmed.ncbi.nlm.nih.gov/41235003/)]

130. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med. Feb 16, 2010;7(2):e1000217. [doi: [10.1371/journal.pmed.1000217](https://doi.org/10.1371/journal.pmed.1000217)] [Medline: [20169112](https://pubmed.ncbi.nlm.nih.gov/20169112/)]

Abbreviations

BAA: Business Associate Agreement
CENTRAL: Cochrane Central Register of Controlled Trials
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Apps and Online Telehealth
FDR: false discovery rate
FM: foundation model
GDPR: General Data Protection Regulation
GenAI: generative AI
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
MI-CLAIM-GEN: Minimum Information about Clinical AI Checklist for Generative Modeling Research
MI-CLEAR-LLM: Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Health Care
PABAK: prevalence-adjusted bias-adjusted Kappa
PHI: protected health information
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
PROSPERO: International Prospective Register of Systematic Reviews
RAG: retrieval-augmented generation
RCT: randomized controlled trial
SAGER: Sex and Gender Equity in Research
SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials
SQUIRE: Standards for Quality Improvement Reporting Excellence
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
TRIPOD-LLM: Transparent reporting of a multivariable model for individual prognosis or diagnosis-LLM

Edited by Stefano Brini; peer-reviewed by Gayathri Surianarayanan, Maria Chatzimina, Yafang Huang; submitted 03.Mar.2026; final revised version received 11.Sep.2026; accepted 14.Sep.2026; published 28.Oct.2026

Please cite as:

Ko JE, Hwang JW, Kim SM, Kwon JY

Reporting Gaps and Challenges in Randomized Trials of Generative AI for Medical Use: Insights From a Systematic Review
J Med Internet Res 2026;28:e94543

URL: <https://www.jmir.org/2026/1/e94543>

doi: [10.2196/94543](https://doi.org/10.2196/94543)

© Jae Eun Ko, Ji Woo Hwang, Sung Min Kim, Ji Yean Kwon. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 28.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.