Protocol

Investigating Placebos and Controls Used in Large Language Model–Based Chatbot Intervention Trials: Protocol for a Methodological Review

Leo Druart[1,2], PT, PhD; Vanda Faria[1,3,4], PhD; Marco Annoni[5], PhD; John Torous[6], MD; Moa Pontén[7,8], PhD; Charlotte Blease[1,6], PhD

[1]Participatory eHealth and Health Data Research Group, Department of Women's and Children's Health, Uppsala University, Uppsala, Uppsala, Sweden

[2]Department of Radiology, Warren Alpert Medical School, Brown University, Providence, RI, United States

[3]Brain and Eye Pain Imaging Lab, Pain and Affective Neuroscience Center, Department of Anesthesiology, Critical Care and Pain Medicine, Boston Children's Hospital and Harvard Medical School, Harvard University, Boston, MA, United States

[4]Comprehensive Pain Center, Faculty of Medicine, University Hospital Carl Gustav Carus, Dresden, Saxony, Germany

[5]Interdepartmental Center for Research Ethics and Integrity, National Research Council, Milan, Lazio, Italy

[6]Department of Psychiatry, Beth Israel Deaconess Medical Center, Division of Digital Psychiatry, Harvard Medical School, Boston, MA, United States

[7]Department of Clinical Neuroscience, Centre for Psychiatry Research, Karolinska Institutet, Stockholm, Stockholm, Sweden

[8]Stockholm Health Care Services, Stockholm, Stockholm, Sweden

Corresponding Author:
Leo Druart, PT, PhD
Participatory eHealth and Health Data Research Group, Department of Women's and Children's Health
Uppsala University
MTC-huset (MTC Building), Dag Hammarskjölds väg 14B, 1 tr
Uppsala, 75237
Sweden
Phone: 46 18 471 00 00
Email: leo.druart@uu.se

Abstract

Background: Large language model (LLM)–based chatbots are rapidly being repurposed as patient-facing digital health tools. Their interactive, adaptive, and seemingly empathic behavior can heighten engagement and expectancy—nonspecific factors that complicate causal inference. Yet, comparator strategies in LLM trials are inconsistently defined and often undermatched (eg, minimal education vs highly engaging chatbots), risking biased effect estimates and poor reproducibility.

Objective: The aim of this study was to systematically identify and categorize the control conditions used in interventional studies of LLM-based, patient-facing digital health interventions and to evaluate their methodological appropriateness. Secondary aims are to describe variability by health domain and study design and to explore whether control type/quality relates to the direction of reported effects.

Methods: This protocol follows PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols) and is registered in PROSPERO. Eligible studies are interventional designs that evaluate LLM-based, patient-facing digital health interventions; any control condition is eligible (including no control, waitlist, treatment-as-usual, attention/education, active comparator, or sham digital control). We will search PubMed, PsycINFO, CENTRAL, CINAHL, and Scopus for records from January 1, 2023, onward. All records will be managed and screened in Rayyan by 2 independent reviewers. Dual, independent data extraction will target study context, intervention details, and control-arm characteristics (typology, rationale, matching to nonspecifics, blinding, reporting). No formal risk-of-bias assessments are planned, as the focus is on meta-research.

Results: At submission, the protocol is registered in PROSPERO and has received no specific funding. Scoping searches are complete; full screening and extraction have not yet commenced.

Conclusions: This review will provide an empirical map of control practices in LLM chatbot trials and guidance for designing better-matched comparators, supporting more valid and interpretable evaluations as LLMs diffuse into patient care.

Trial Registration: PROSPERO CRD420251246148; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251246148

International Registered Report Identifier (IRRID): PRR1-10.2196/90507

(JMIR Res Protoc 2026;15:e90507) doi: 10.2196/90507

KEYWORDS

large language models; chatbots; digital health; control conditions; methodological review

Introduction

Background

Since ChatGPT's launch in November 2022, interest in patient-facing conversational agents whose replies are generated by large language models (LLM), in other words, LLM-based chatbots, has surged, with growing attention to their potential clinical applications. Just a year after its launch, already 1 in 5 primary care physicians in the United Kingdom reported using it in their professional activities [1]. While the application of LLMs in health care presents significant opportunities to improve health care, for example, by diminishing the administrative burden for clinicians while also helping patients through its accessibility and availability, they also invite significant challenges that warrant consideration [2,3].

Now, patients are using these chatbots for their health care needs. For example, by early 2025, 10% of Australians had already used ChatGPT for health-related questions, with 61% asking questions about taking action that typically require clinical advice [4]. Of the participants who had not yet used ChatGPT for a health-related question, 40% reported that they would do so within the next 6 months. More striking, in October 2025, OpenAI announced that, every week, 1.2 million users disclosed thinking about taking their own lives with ChatGPT [5]. It is clear that these tools are being repurposed for lifestyle counselling or mental health support. Studies demonstrate that these chatbots can simulate empathy, personalize responses, and sustain meaningful dialogue, potentially amplifying expectancy and engagement effects beyond those seen in rule-based apps [6].

One key challenge, however, lies in measuring their efficacy and determining whether current methodological tools are adapted to assess these new interventions. While rigorous trials are a prerequisite for introducing new drugs to patients, LLM tools have not been subjected to similar scrutiny. Ideally, an LLM should only be used as a health intervention after safety and effectiveness have been demonstrated [7].

Such rapid dissemination raises concerns that deployment may have outpaced the development of appropriate evaluation practices. Like other nonpharmacological interventions such as surgery, physiotherapy, or psychotherapy [8], LLMs, and more generally digital interventions, pose unique challenges for evaluation when compared to drugs. Their effects often rely on multiple components, and they are, as interventions, highly adaptive. Moreover, they inherently depend on user engagement, interface design, and personalization, all of which could potentially influence outcomes independently of any hypothesized active ingredient. As the field evolves, these methodological specificities will need greater methodological attention [9]. Specifically, methodological attention should focus on designing appropriate control groups for evaluating these tools. Controls are necessary to sift out any confounding factors and accurately assess the efficacy of LLM interventions [10]. Yet, adequate controls are challenging to design and implement [11]. These considerations are especially salient for unsupervised LLMs whose conversational, adaptive, and apparently empathic responses may heighten engagement and expectancy effects [6,12,13]. In this context, it is important to distinguish between placebo responses—any observed change following a placebo intervention, and placebo effects, proper, which refer specifically to the component of that response attributable to placebo mechanisms such as expectancy or conditioning [11,14].

This is why the choice of control condition is fundamental to assess the efficacy of these tools—a consideration that has not only scientific but also ethical implications [12]. Foundational guidance for nonpharmacological trials emphasizes that controls should be selected to answer a specific scientific question while accounting for nonspecific factors such as time, attention, therapeutic alliance, and expectancy effects [15,16]. In LLM trials, an adequate control may need to match not only time and attention but also the look-and-feel and level of interactive engagement (eg, notifications, chat, interface aesthetics) to approximate a placebo control that isolates the hypothesized mechanism of action [17]. Without such methodological matching, apparent benefits may be inflated (if controls are minimally engaging) or genuine effects obscured (if controls inadvertently contain active elements), limiting the validity of the findings.

Recent scholarship has begun to map the challenges in digital health more generally. Goldberg and colleagues [17] proposed a typology of 11 control condition types used in mobile health randomized trials, along with a decision framework and a gradation of comparison strength, illuminating how different control choices afford different inferences. Methodological work has urged for comparators that better approximate what some researchers have dubbed "active placebo" characteristics in digital trials—mirroring the active arm's interactivity and engagement while withholding the hypothesized mechanism [17,18]. A systematic review of digital shams in digital therapeutics highlighted practical and conceptual challenges—among them, inconsistently defined controls, inadequate matching on engagement, and limited reporting [18].

These contributions underscore that controls are heterogeneous, variably justified, and unevenly reported in digital intervention trials and that stronger standards are needed to ensure valid inference and comparability across studies. However, this has yet to be done for LLM-based interventions, which are increasingly gaining popularity.

Scoping reviews show that evaluation methods for LLMs are lacking in high-quality effectiveness research. The evaluation of LLMs in coaching and lifestyle has been suggested to present a lack of methodological standardization and a general trend toward less rigorous evaluation designs [19]. Similar findings were reported for LLMs used in mental health [20]. Moreover, effectiveness has been defined subjectively in over two-thirds of studies evaluating LLMs to provide health advice [21]. These findings highlight that the need for improved control designs is particularly acute for LLM, given their widespread use has advanced faster than their systematic evaluation.

In summary, the exponential growth of LLMs used as patient-facing digital health tools and the unique methodological challenges they present create a critical empirical gap: we lack a consolidated mapping of the control conditions actually used across these interventional studies of LLM-based chatbots. We also need to better understand how consistently these controls are justified and described and the extent to which they match nonspecific factors that are especially salient for conversational/artificial intelligence (AI) systems. Moreover, while guidance documents (eg, CONSORT-EHEALTH [Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth]) have improved reporting expectations, they do not extend to control adaptations such as which control to use for a given scientific aim, and evidence suggests that control selection and labeling (eg, treatment as usual, education, waitlist) remain inconsistent across the literature [17,22]. A methodological synthesis focused specifically on LLM-based interventions, where conversationality and perceived intelligence may intensify nonspecific effects, can provide the field with a clearer description of practices, highlight reporting deficits, and inform design decisions in upcoming trials.

Accordingly, this protocol describes a methodological review that aims to answer the question: what type of control conditions are used in interventional studies of LLM-based, patient-facing digital health interventions, and how methodologically appropriate are these controls?

Objectives

This methodological review aims to systematically identify and categorize the control conditions used in interventional studies of LLM-based chatbot interventions and to evaluate their methodological appropriateness—specifically, whether a clear rationale is provided, whether placebo response components are adequately matched, and whether risks of blinding failure or contamination are addressed and reported transparently. We will also describe how control design and reporting vary across 3 health domains (mental health, chronic disease, lifestyle changes) and study designs and explore whether the type or quality of control relates to the direction of reported intervention effects (positive, null, or negative findings).

The population consists of individuals in any health domain who are intended users of an AI-based digital health tool. The intervention is an LLM-based, patient-facing digital modality, and the comparator may be any control condition, including no control, a waitlist, treatment-as-usual, attention/education, an active comparator, or another type of intervention control. The outcomes of the review include the methodological description, classification, and appraisal of control arms rather than clinical effectiveness; notwithstanding, the quality of the former will be relevant to measuring the latter.

Methods

Study Design

This is a protocol for a methodological (meta-research) review reported in accordance with PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols) [23]. Where relevant, general methods are informed by the Cochrane Handbook [24]. Methodological reviews identify inconsistencies, gaps, and sources of bias in the design, conduct, and reporting of research. They are particularly useful when an area of research is rapidly evolving and lacks standardization, allowing researchers to map current practices and evaluate their adequacy [25]. Given the heterogeneity and emerging nature of control conditions in LLM-based digital health interventions, this approach enables a rigorous assessment of whether existing trials use appropriate comparators and adhere to methodological best practices.

Eligibility Criteria

We will include interventional studies evaluating LLM-based health interventions, including randomized controlled trials, nonrandomized trials, quasi-experimental studies, and single-group trials (pilot/feasibility studies). Published study protocols may be considered. We will exclude observational designs (cohort, cross-sectional, case-control), case reports, editorials, reviews and commentaries, and non–peer-reviewed publications.

Regarding the population, we will include studies with individuals of any age/sex/health status who are intended users of the intervention with no restriction by clinical domain. Subpopulation analyses will be conducted for mental health, lifestyle change, or chronic disease management interventions.

The interventions concerned will be LLM-based chatbots used as patient-facing digital health interventions in which LLM constitutes the active ingredient (eg, mobile apps, web platforms, chatbots/virtual agents, generative-AI tools). We will exclude studies where the digital component serves only as a delivery vehicle (eg, videoconferencing for standard therapy) and interventions that are not directly patient-facing. An LLM will be considered the active ingredient when it generates patient-facing linguistic output intended to contribute to the intervention experience, including dialogue, explanation, summarization, or rephrasing. This reflects theoretical and empirical work suggesting that AI-generated language itself may influence expectancy, meaning-making, or engagement [11,26]. By contrast, LLM uses limited to back-end processing, or clinician-facing functions will be excluded.

Regarding controls, any control condition is eligible, including none (eg, no-treatment), waitlist, treatment-as-usual, active comparators, attention controls, placebo/sham digital controls, or standard care. Lastly, clinical outcomes will not be the eligibility criteria in this review. As a methodological review, our primary outcomes concern the characteristics of the control arms used in the included studies (see “Data items”). Included publications will be in English and published from January 2023 onward to reflect the adoption and investigation of LLMs. This cutoff was selected because it marks the point at which patient-facing, conversational LLM-based chatbots became both technically plausible and widely deployed at scale. Earlier generative or GPT-based systems were typically experimental, constrained, or not deployed as adaptive patient-facing interventions, and are therefore methodologically distinct from the intervention class examined in this review. An additional sense of urgency into the effectiveness of these tools is required because LLM-based chatbots are reported being adopted at scale by consumers and patients. For example, OpenAI reported in 2025 that 1.2 million adults every month disclosed suicidal thoughts to ChatGPT [5] and a recent UK study found that 37% of adults had used the tools for emotional or mental health support [27].

Search Strategy

We will search 5 major bibliographic databases: PubMed, PsycINFO via EBSCO, CENTRAL via Cochrane Library, CINAHL via EBSCO, and Scopus. We will screen references of included studies and relevant reviews and consider backward and forward citation tracking. The coverage period will begin on January 1, 2023, and the search will be performed by the Uppsala University librarian team on the project (Görel Sundström and Emma Kristina Carlsson). Given the rapid pace of development in this field, we will conduct an updated search and incorporate any newly identified studies before submitting the manuscript or during the peer-review process.

The final database-specific strategies will be designed in collaboration with study librarians. The search strategies will be peer-reviewed in accordance with the PRESS 2015 framework [28]. The search strategy will be designed for PubMed using Medical Subject Headings (MeSH) and then adapted to the other databases and their thesauri. Terms will also be searched for in the title/abstract of publications. The PubMed search string is available in Multimedia Appendix 1.

Data Management and Selection Process

Records will be managed in Rayyan for deduplication and screening [29]. Screening will proceed in two stages, that is, (1) titles/abstracts and (2) full texts, conducted independently by two reviewers through the Rayyan software. Disagreements will be resolved by discussion or by a third reviewer. We will report a PRISMA 2020 flow diagram for identification, screening, eligibility, and inclusion [30]. Prior to formal screening, we will pilot test on a smaller sample of studies. Any disagreements at this stage will be discussed and help refine the screening process to ensure consistent interpretation.

Data Extraction

At least two reviewers will independently extract data by using an extraction form. Disagreements will be resolved through discussion or by a third reviewer. Where intervention or control details are unclear, we will attempt to contact the author (via one initial email and one reminder). We will record when information is not reported or unclear. Extracted data are presented in Table 1.

Table 1. Planned extracted data.

Type of data	Data
Study identification	• First author name • Year of publication • Country of study
Study design and sample	• Study design (options: randomized controlled trial, nonrandomized trial, single group) • Sample size • Participant characteristics (age, gender, clinical condition, self-report or clinical diagnosis, verification of self-reported data) • Recruitment (online, in-person) • Follow-up characteristics (length of study, compensation or incentives) • Research aim (copy research aim) • Requires a control group? (deduced from research aim, possible options: yes, no, unsure)
Intervention characteristics	• Name/description of large language model intervention • Mode of delivery (eg, mobile app, web-based content) • Target condition or domain • Duration, frequency, and intensity of the intervention • Supposed mechanism of action explained (yes/no) • Supposed mechanism of action (detail in free text)
Comparator characteristics	• Type of control condition • Type of control condition (describe in free text) • Justification for choice of control explained (yes/no) • Justification for choice of control (detail in free text) • Blinding status (participants, providers, assessors, analysts)
Methodological appropriateness	• Given the study design, research aim, need for control, intervention description, supposed mechanism of action and type of control condition, is the control group appropriate? (Yes/No) • Comment on the previous answer • Are placebo responses controlled for adequately? • Are placebo effects controlled for adequately?
Outcomes and results	• Conclusion on primary outcome (positive finding, null finding, unclear)

Outcomes

The primary outcome of this review is the type and frequency of control conditions used. The absence of a comparator will be coded as a distinct control category (no control). Control-arm appropriateness will be evaluated relative to the stated research aim; where a comparator is required to support causal inference but absent, the study will be classified as not appropriately controlled.

Secondary outcomes will include the adequacy/appropriateness of controls with respect to the placebo response and theoretical justification; the reporting quality of control descriptions and justifications. Control-arm appropriateness will be evaluated using a predefined set of extraction items (Table 1: items in methodological appropriateness), which are considered jointly as a qualitative decision framework. These items assess whether a control condition is present, justified, aligned with the study aim, and designed to address relevant placebo response components. Appropriateness will be classified categorically rather than scored numerically, as different control choices afford different inferential aims rather than representing a linear quality continuum. Where judgments require interpretation, disagreements will be resolved through collegial discussion among authors with complementary expertise in placebo research, clinical trial methodology, and digital health interventions [11,31]. For Table 1 items regarding the methodological appropriateness of accounting for placebo responses and effects, we will consider that placebo responses refer to observed changes following intervention exposure, whereas placebo effects proper denote changes attributable to mechanisms such as expectancy or conditioning [11,14].

An exploratory outcome is to assess whether the type and quality of the control condition are associated with differences in the reported effectiveness of digital interventions, by comparing success proportions (see Table 1) across control categories.

Risk of Bias in Individual Studies

We will not conduct a formal risk-of-bias assessment because the review does not synthesize treatment effects; instead, it characterizes control-arm design and reporting (ie, a meta-research focus). We will, however, extract bias-relevant features (eg, blinding, contamination, missing data statements) to inform the appraisal of control adequacy.

Assessment of publication bias or selective reporting at the meta-analytic level is not planned, as we do not pool effect sizes. GRADE (Grading of Recommendations Assessment, Development and Evaluation) is not applicable because we are not grading the certainty of an effect estimate [32].

Data Synthesis

We will use descriptive statistics (counts, proportions, medians/IQRs where relevant) to summarize control-type frequencies, presence/absence of justifications, and reporting completeness. Stratified summaries will be presented by health domain (mental health, lifestyle changes, chronic disease management). For studies with more than two arms, we will treat each LLM-vs-control comparison in a multi-arm study as a separate comparison unit for exploratory analyses of outcome direction by control type.

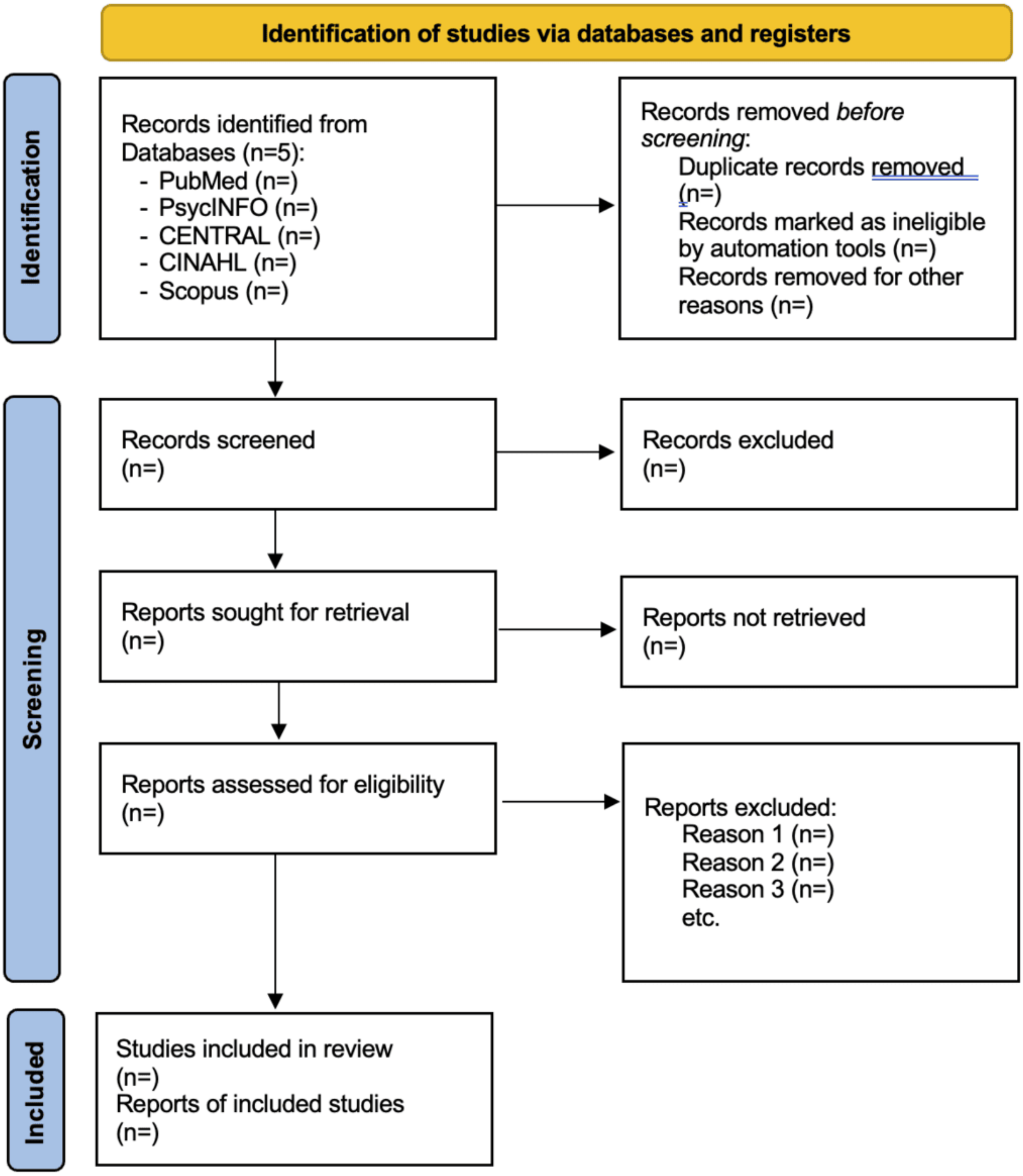
Amendments

Any protocol amendments (eg, eligibility refinements, added subgroup) will be documented, dated, and justified within the PROSPERO record and reported in the final manuscript [23].

Results

As of submission, the protocol has been registered in PROSPERO (CRD420251246148). The database-specific search strategies have been pretested, and preliminary scoping searches have been conducted to estimate the volume and refine the eligibility criteria. No study data have been synthesized yet; this manuscript details the a priori methods that will guide the review as shown in Figure 1.

Figure 1. Anticipated PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) 2020 flow diagram for the planned study selection process.


Identification of studies via databases and registers
Identification
Screening
Included
Records identified from Databases (n=5):
- PubMed (n=)
- PsycINFO (n=)
- CENTRAL (n=)
- CINAHL (n=)
- Scopus (n=)
Records removed *before screening*:
Duplicate records removed (n=)
Records marked as ineligible by automation tools (n=)
Records removed for other reasons (n=)
Records screened (n=)
Records excluded (n=)
Reports sought for retrieval (n=)
Reports not retrieved (n=)
Reports assessed for eligibility (n=)
Reports excluded:
Reason 1 (n=)
Reason 2 (n=)
Reason 3 (n=)
etc.
Studies included in review (n=)
Reports of included studies (n=)

Discussion

Principal Results

This review will provide researchers, clinicians, and policymakers with a systematic assessment of the control conditions used to measure the efficacy of LLM-based interventions and their adequacy. This will enable a more comprehensive evaluation of the current evidence regarding the effectiveness of LLM interventions, allowing health care providers, policymakers, and, most importantly, patients to make informed decisions when considering the use of LLMs in health care.

We hypothesize that many LLM-based health trials employ control conditions that are variably defined, underjustified, and insufficiently matched for nonspecific effects, reflecting the specific challenges that evaluating digital interventions pose, yet potentially biasing inferences about the efficacy of the intervention. The outcome of this research will be highly informative for all stakeholders and will allow researchers to design more optimal control conditions in the future.

Limitations

A key limitation of this review is that the evidence base for LLM-based digital health interventions is rapidly evolving, which restricts the strength of any conclusions about optimal control conditions. This challenge is further accentuated by the inherent delay between the emergence of new technologies, the conduct of research, and the publication of findings. Nonetheless, adoption of these systems is already occurring at scale, and the ethical risks associated with inadequate or poorly designed controls are substantial. In this context, we argue that providing an early methodological assessment is preferable to delaying synthesis until the field matures, as timely guidance may help prevent avoidable methodological weaknesses from becoming entrenched in future research.

Comparison With Prior Work

There has yet to be any systematic mapping of the design and use of control conditions in LLM trials. However, research has been performed for digital interventions more broadly, allowing researchers to adapt reporting guidelines to the specifics of eHealth interventions through the CONSORT-EHEALTH, emphasizing the need to specifically describe components such as usability, dose/exposure, and usage metrics [22].

A previous review provided an initial examination of how control conditions were used in digital therapeutics; however, this work preceded the advent of LLMs in health care [18]. Goldberg and colleagues [17] offered a precise mapping of the typology of control conditions that were used in mobile health interventions, showing a clear need for improvement.

Conclusions

This protocol outlines a timely and rigorous plan for a methodological synthesis of control conditions used in LLM-based health trials. By systematically classifying control arms, assessing their appropriateness, and evaluating the quality of reporting, the review will provide an empirical foundation for improved comparator design in future LLM health research. The findings are expected to inform investigators and reviewers in the rapidly evolving field of LLM-based chatbots and generative AI, where standards for comparators are still being developed.

Acknowledgments

The authors would like to thank Emma Kristina Carlsson and Görel Sundström of Uppsala University Library for their assistance in developing the search strategy for this research protocol. The authors declare the use of generative artificial intelligence (GAI) in the manuscript preparation process. According to the GAIDeT (GAI Delegation Taxonomy, 2025), the following tasks were, in part, delegated to GAI tools under full human supervision: (1) proofreading and editing and (2) publication support (fitting word counts). The GAI tool used was ChatGPT5.2. The responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

This article is a research protocol and does not include any data. Data resulting from this research will be made available upon publication of the results.

Conflicts of Interest

CB is an Associate Editor of *JMIR Mental Health* and JT is Editor-in-Chief of *JMIR Mental Health*. This manuscript was handled by an academic editor with no conflicts of interest. CB received funding from Google DeepMind in 2026 for independent ethics work, which did not influence the content of this manuscript.

Multimedia Appendix 1

PubMed search strategy.
[DOCX File , 14 KB-Multimedia Appendix 1]

References

1. Blease CR, Locher C, Gaab J, Hägglund M, Mandl KD. Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ Health Care Inform. Sep 17, 2024;31(1):e101102. [FREE Full text] [doi: 10.1136/bmjhci-2024-101102] [Medline: 39288998]
2. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health. Nov 2023;26(1):e300884. [FREE Full text] [doi: 10.1136/bmjment-2023-300884] [Medline: 37949485]
3. Faria V, Goturi N, Dynak A, Talbert C, Pondelis N, Annoni M, et al. Triadic relations in healthcare: surveying physicians' perspectives on generative AI integration and its role on empathy, the placebo effect and patient care. Front Psychol. 2025;16:1612215. [FREE Full text] [doi: 10.3389/fpsyg.2025.1612215] [Medline: 41312268]

4. Ayre J, Cvejic E, McCaffery K. Use of ChatGPT to obtain health information in Australia, 2024: insights from a nationally representative survey. Med J Aust. Mar 03, 2025;222(4):210-212. [FREE Full text] [doi: 10.5694/mja2.52598] [Medline: 39901778]
5. 2025. Amélioration des réponses de ChatGPT lors de conversations sensibles. URL: https://openai.com/fr-FR/index/strengthening-chatgpt-responses-in-sensitive-conversations/ [accessed 2025-12-29]
6. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. Jun 01, 2023;183(6):589-596. [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
7. Kohane IS. Injecting artificial intelligence into medicine. NEJM AI. Jan 2024;1(1):1. [doi: 10.1056/aie2300197]
8. Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P, CONSORT NPT Group. CONSORT statement for randomized trials of nonpharmacologic treatments: a 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. Ann Intern Med. Jul 04, 2017;167(1):40-47. [FREE Full text] [doi: 10.7326/M17-0046] [Medline: 28630973]
9. van Mierlo T. From efficacy to scale: addressing digital health's original sin. J Med Internet Res. Oct 27, 2025;27:e85878. [FREE Full text] [doi: 10.2196/85878] [Medline: 41143679]
10. Torous J, Firth J. The digital placebo effect: mobile mental health meets clinical psychiatry. Lancet Psychiatry. Feb 2016;3(2):100-102. [doi: 10.1016/S2215-0366(15)00565-9] [Medline: 26851322]
11. Druart A, Blease C. Should digital health trials be placebo controlled? JME Practical Bioethics Institute of Medical Ethics. Dec 10, 2025;1:1. [doi: 10.1136/jmepb-2025-000027]
12. Blease C. Placebo, nocebo, and machine learning: how generative AI could shape patient perception in mental health care. JMIR Ment Health. Aug 15, 2025;12:e78663. [FREE Full text] [doi: 10.2196/78663] [Medline: 40815809]
13. Kharko A, McMillan B, Hagström J, Muli I, Davidge G, Hägglund M, et al. Generative artificial intelligence writing open notes: a mixed methods assessment of the functionality of GPT 3.5 and GPT 4.0. Digit Health. 2024;10:20552076241291384. [FREE Full text] [doi: 10.1177/20552076241291384] [Medline: 39493632]
14. Evers AWM, Colloca L, Blease C, Annoni M, Atlas LY, Benedetti F, et al. Implications of placebo and nocebo effects for clinical practice: expert consensus. Psychother Psychosom. 2018;87(4):204-210. [FREE Full text] [doi: 10.1159/000490354] [Medline: 29895014]
15. Locher C, Gaab J, Blease C. When a placebo is not a placebo: problems and solutions to the gold standard in psychotherapy research. Front Psychol. 2018;9:2317. [FREE Full text] [doi: 10.3389/fpsyg.2018.02317] [Medline: 30542310]
16. Blease C. Psychotherapy and placebos: manifesto for conceptual clarity. Front Psychiatry. 2018;9:379. [FREE Full text] [doi: 10.3389/fpsyt.2018.00379] [Medline: 30177892]
17. Goldberg SB, Sun S, Carlbring P, Torous J. Selecting and describing control conditions in mobile health randomized controlled trials: a proposed typology. NPJ Digit Med. Sep 30, 2023;6(1):181. [FREE Full text] [doi: 10.1038/s41746-023-00923-7] [Medline: 37775522]
18. Bos E, Preller KH, Kaur G, Malhotra P, Kharawala S, Motti D. Challenges with the use of digital sham: systematic review and recommendations. J Med Internet Res. Oct 24, 2023;25:e44764. [FREE Full text] [doi: 10.2196/44764] [Medline: 37874638]
19. Lai X, Lai Y, Chen J, Huang S, Gao Q, Huang C. Evaluation strategies for large language model-based models in exercise and health coaching: scoping review. J Med Internet Res. Oct 14, 2025;27:e79217. [FREE Full text] [doi: 10.2196/79217] [Medline: 41086432]
20. Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, et al. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med. Apr 30, 2025;8(1):230. [FREE Full text] [doi: 10.1038/s41746-025-01611-4] [Medline: 40307331]
21. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open. Feb 03, 2025;8(2):e2457879. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
22. Eysenbach G. CONSORT-EHEALTH: implementation of a checklist for authors and editors to improve reporting of web-based and mobile randomized controlled trials. Stud Health Technol Inform. 2013;192:657-661. [Medline: 23920638]
23. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. Jan 02, 2015;350:g7647. [FREE Full text] [doi: 10.1136/bmj.g7647] [Medline: 25555855]
24. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane Handbook for Systematic Reviews of Interventions. 1st ed. Wiley. USA. Wiley; 2019.
25. Mbuagbaw L, Lawson DO, Puljak L, Allison DB, Thabane L. A tutorial on methodological studies: the what, when, how and why. BMC Med Res Methodol. Sep 07, 2020;20(1):226. [FREE Full text] [doi: 10.1186/s12874-020-01107-7] [Medline: 32894052]
26. Song E, Bernstein M, Lay P, Druart L, Dibble E, Lourenco A, et al. Accessing AI mammography reports impacts patient follow-up behaviors: the unintended consequences of including AI in patient portals. NPJ Digit Med. Dec 03, 2025;8(1):761. [FREE Full text] [doi: 10.1038/s41746-025-02201-0] [Medline: 41339742]

27. Orpwood G. Over one in three using AI Chatbots for mental health support, as charity calls for urgent safeguards. Mental Health UK. URL: https://mentalhealth-uk.org/blog/over-one-in-three-using-ai-chatbots-for-mental-health-support-as-charity-calls-for-urgent-safeguards/ [accessed 2026-02-24]
28. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol. Jul 2016;75:40-46. [FREE Full text] [doi: 10.1016/j.jclinepi.2016.01.021] [Medline: 27005575]
29. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. Dec 05, 2016;5(1):210. [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
30. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. Mar 29, 2021;372:n160. [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
31. Blease C, Druart L. Difficulty is not an ethical exemption: placebo controls and epistemic responsibility in digital therapeutics. JME Practical Bioethics. Feb 12, 2026;2(1):e000129-e000141. [doi: 10.1136/jmepb-2026-000129]
32. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. Apr 26, 2008;336(7650):924-926. [FREE Full text] [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]

Abbreviations

AI: artificial intelligence
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
GRADE: Grading of Recommendations Assessment, Development and Evaluation
LLM: large language model
MeSH: Medical Subject Headings
PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analysis
PRISMA-P: Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols

Edited by J Sarvestan; submitted 29.Dec.2025; peer-reviewed by L He, H Su; comments to author 04.Feb.2026; revised version received 24.Feb.2026; accepted 03.Mar.2026; published 17.Mar.2026

Please cite as:
Druart L, Faria V, Annoni M, Torous J, Pontén M, Blease C
Investigating Placebos and Controls Used in Large Language Model–Based Chatbot Intervention Trials: Protocol for a Methodological Review
JMIR Res Protoc 2026;15:e90507
URL: https://www.researchprotocols.org/2026/1/e90507
doi: 10.2196/90507
PMID: