Protocol

Exploring Bias in Medical Applications of Large Language Models: Protocol for a Systematic Review

Shweta Rao[1]; Carol Boutrous[1]; Andrey Kormilitzin[2], BSc, MSc, PhD; Xin You Tai[3], DPhil, MRCP, MBBS, MSc; Lewis Hotchkiss[4], BSc; Cen Cong[5], BEng, MA, PhD; Edward Meinert[6], MA, MSc, MBA, MPA, MPH, PhD; Huizhi Liang[7], PhD; Hang Dong[8], BSc, MSc, PhD; Judith R Harrison[6], MBChB, MRCPsych, PhD

[1]School of Medicine, Newcastle University, Newcastle, AK, United Kingdom
[2]Department of Psychiatry, University of Oxford, Headington, Oxford, England, United Kingdom
[3]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, England, United Kingdom
[4]Dementias Platform UK (DPUK), Swansea University, Swansea, Wales, United Kingdom
[5]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle, England, United Kingdom
[6]Translational and Clinical Research Institute 3rd Floor, Biomedical Research Building Campus for Ageing and Vitality, Newcastle, England, United Kingdom
[7]School of Computing, Newcastle University, Newcastle upon Tyne, United Kingdom
[8]Department of Computer Science, University of Exeter, Exeter, United Kingdom

Corresponding Author:

Judith R Harrison, MBChB, MRCPsych, PhD
Translational and Clinical Research Institute 3rd Floor
Biomedical Research Building Campus for Ageing and Vitality
Westgate Road, Newcastle upon Tyne NE4 5PL
Newcastle, England
United Kingdom
Phone: 44 7756339941
Email: Judith.Harrison@newcastle.ac.uk

Abstract

Background: Large language models (LLMs) are increasingly applied in health care for clinical decision-making, education, and patient communication. However, bias in LLM outputs may exacerbate health care disparities and compromise trust. Despite rapid adoption, there is limited synthesis of how bias is identified, measured, and mitigated in medical applications of LLMs.

Objective: This systematic review aims to evaluate how bias is detected, measured, and mitigated in health care applications of LLMs and to identify methodological trends and gaps in the current literature.

Methods: We will conduct a systematic review of studies evaluating bias, fairness, or subgroup performance in medical applications of LLMs. Searches will be conducted across Embase, MEDLINE, PsycINFO, PubMed, ACL Anthology, ACM Digital Library, arXiv, medRxiv, and bioRxiv for studies published from 2017 onward. The review uses a staged design. In phase 1, records identified in the original June 2025 search were screened and assessed manually using predefined eligibility criteria. This fully manually reviewed dataset will serve as a reference standard for validating the review workflow. In phase 2, an updated June 2026 search will be screened using a prespecified LLM-assisted workflow. Two independent LLMs will apply the same eligibility criteria used in manual screening, with studies marked as potentially relevant or uncertain by either model progressing to further assessment. Final inclusion decisions will be made by at least 2 human reviewers. The performance of the LLM-assisted workflow will be evaluated against the manually reviewed reference dataset using sensitivity, specificity, precision, negative predictive value, inclusion agreement, and Cohen κ. Data will be extracted using a standardized framework covering study characteristics, model details, health care use case, bias type, bias assessment methods, mitigation strategies, transparency, generalizability, and ethical or regulatory framing. Findings will be synthesized narratively.

Results: The phase 1 search, conducted on June 11, 2025, yielded 15,976 records after deduplication and has undergone complete manual screening. The phase 2 search, conducted on June 8, 2026, increased the total number of records to 41,143. Screening of the updated dataset using the LLM-assisted workflow is ongoing. The completed review will report included study characteristics, approaches to bias detection and mitigation, and validation metrics for the LLM-assisted workflow. The

review is expected to be submitted for publication in late November 2026, with publication anticipated thereafter subject to the journal's peer-review and editorial process.

Conclusions: This review aims to provide a comprehensive synthesis of current approaches to bias detection and mitigation in medical LLMs, highlighting methodological strengths, limitations, and areas for future research. The findings aim to inform the development of more equitable and transparent AI systems in health care.

Trial Registration: PROSPERO CRD420250638943; https://www.crd.york.ac.uk/PROSPERO/view/CRD420250638943 and OSF Registries osf.io/szjhc; https://osf.io/szjhc/

International Registered Report Identifier (IRRID): DERR1-10.2196/91348

JMIR Res Protoc 2026;15:e91348; doi: 10.2196/91348

Keywords: large language model; LLM; bias; health care; medicine

Introduction

Large language models (LLMs) are advanced transformer-based AI systems trained on massive volumes of text data, enabling them to generate fluent, contextually appropriate language. Recent innovations in LLM architectures, including Generative Pretrained Transformer [1], Bidirectional Encoder Representations from Transformers [2], Pathways Language Model [3], and Large Language Model Meta AI [4], have significantly expanded the scope of these models. In medicine, LLMs are increasingly applied to tasks such as clinical decision support [5], automated documentation [6], patient communication [7], and biomedical literature analysis [8], offering the potential to improve efficiency, accessibility, and patient engagement.

Unlike traditional rule-based systems, LLMs can synthesize unstructured medical data, respond dynamically in conversational contexts, and assist with real-time medical information retrieval [9]. Their capacity to model nuanced language use makes them promising tools for augmenting health care delivery and education. However, this flexibility also introduces substantial risk: LLMs are inherently probabilistic and operate through high-dimensional latent representations and self-attention mechanisms [10]. As a result, their outputs are nondeterministic, opaque, and potentially inconsistent [11], posing challenges for transparency, verifiability, and safety in clinical settings [6,12,13].

Bias in LLMs refers to systematic errors in output that arise from imbalances in training data, annotation processes, or architectural constraints [14]. In this review, bias refers specifically to systematic differences in model performance or outputs across demographic or clinically relevant subgroups. This is conceptually distinct from related phenomena such as hallucination, which refers to the generation of factually incorrect or unsupported information [15], and from broader considerations relating to AI safety, accountability, and transparency. While these issues may co-occur in practice, this review focuses on bias as it relates to systematic inequities in outputs.

In health care, such biases can manifest as disparities in diagnostic accuracy, therapeutic recommendations, or patient communication. For instance, if training data overrepresent certain populations or reflect historically biased clinical guidelines, LLM outputs may reinforce these inequities. Notable examples include race-based estimations in kidney function [16] or gender disparities in cardiovascular diagnosis and treatment [17], biases that LLMs may inherit and perpetuate if not explicitly addressed.

The complexity of LLMs compounds the difficulty of bias detection and mitigation. Because these models generate outputs without a fixed decision pathway, biases may appear subtly and variably, eluding traditional evaluation metrics [18]. Techniques such as dataset audits, counterfactual fairness testing, and adversarial evaluation have emerged to assess and quantify bias, yet no unified framework exists for their systematic application in health care LLMs [18-20]. Moreover, while some mitigation strategies, such as data augmentation, algorithmic debiasing, and retrieval-augmented generation (RAG) [21-23], have been proposed, their real-world effectiveness remains uncertain, and in some contexts, these approaches may also introduce or exacerbate bias.

Recent systematic reviews have begun to examine aspects of bias and fairness in medical LLMs, including demographic disparities in model performance [24]. However, these studies primarily focus on identifying and describing bias rather than systematically evaluating the methodologies used to detect and address it. Other reviews have either focused on classical machine learning or deep learning in medical imaging and disease diagnosis [25-32] or have surveyed LLM applications in health care without a primary focus on fairness or bias [33]. As the integration of LLMs into health care accelerates, there is a growing need to systematically evaluate how bias is identified, measured, and mitigated. This review aims to address this gap by synthesizing current approaches across medical applications. It focuses specifically on LLM uses with direct or potential near-term clinical relevance to ensure that findings are applicable to real-world health care settings.

The rapid growth of the literature on medical LLMs presents challenges for evidence synthesis, particularly where large numbers of records require screening and assessment. Recent studies have explored the use of LLMs to support systematic reviews, demonstrating improved efficiency while maintaining high agreement with human reviewers [34]. Such approaches may provide a scalable means of managing increasing volumes of evidence while preserving methodological rigor through appropriate validation procedures.

Methods

Objectives

This systematic review aims to identify and evaluate how bias is detected, measured, and mitigated in applications of LLMs within medical contexts. Specifically, it will

1. map the medical use cases of LLMs and determine in which domains bias has been evaluated (eg, clinical decision support, documentation, patient communication, education, or research).
2. assess the methodologies employed for bias detection and measurement, including dataset audits, counterfactual fairness testing, adversarial evaluation, and explainability techniques.
3. catalog the bias mitigation strategies reported in the literature, such as data augmentation, algorithmic debiasing, RAG, and post hoc correction techniques.
4. synthesize the evidence into preliminary best practice guidance, highlighting effective approaches for bias assessment and mitigation to inform future research, development, and regulatory oversight of LLMs in health care.

Research Questions

To ensure a structured approach to defining the scope of this systematic review, we adopted the Population-Concept-Context (PCC) framework [35] for systematic reviews that aim to map methodologies rather than evaluate interventions. In this review:

- Population (P): LLMs applied to health care.
- Concept (C): Bias detection methods and mitigation strategies.
- Context (C): Clinical, educational, and research applications of LLMs in medicine.

Based on this framework, our review addresses the following research questions:

- What medical applications use LLMs, and in which of these has bias been assessed?
- What methods have been used to detect and measure bias in medical LLMs, including those related to demographic disparities (eg, race, gender, age), clinical misrepresentation, or dataset imbalances?
- What mitigation strategies have been implemented to reduce bias in medical LLMs?
- To what extent do current studies report on the effectiveness or limitations of these bias mitigation approaches?
- Based on existing evidence, what best practices can be recommended for future research on bias assessment and mitigation in medical LLMs?

Eligibility Criteria

For the purposes of this review, LLMs are broadly defined as transformer-based natural language processing models applied to health care, including encoder-only, decoder-only, and encoder-decoder architectures. Encoder-only models (eg, BERT variants) were included because they remain widely used in health care fairness and subgroup-performance research. Multimodal or foundation models incorporating vision-language or speech-language capabilities will be included only where language generation or language-based clinical decision support forms a substantial component of the evaluated system.

The following inclusion criteria were established to ensure that selected studies directly investigate bias in medical LLM applications and provide empirical findings on detection or mitigation strategies:

1. Studies must involve an LLM or foundation language model applied to a medical or health care context and assess bias, fairness, or subgroup performance in model outputs.
2. Eligible studies include experimental, simulation, benchmarking, audit, or evaluation studies using real-world or synthetic datasets, provided they assess bias, fairness, or subgroup performance within a clinically relevant health care context.
3. The study must be published in English.
4. Papers published from 2017 onward, corresponding to the publication of the Transformer architecture [36], which forms the foundational basis of contemporary LLMs.
5. Studies examining training or fine-tuning will be included only if they evaluate model outputs, performance, or bias in a clinically relevant context.
6. Gray literature (eg, preprints, conference proceedings, technical or institutional reports) will be included if sufficient methodological detail, clear data sources, and reported findings relevant to bias or fairness are provided.

The following exclusion criteria were defined to remove studies that do not focus on bias in medical LLMs, lack empirical evaluation, or investigate nonmedical AI applications:

1. Studies focusing solely on machine learning models or algorithms other than LLMs.
2. Reviews, editorials, letters, commentaries, opinion pieces, magazine articles, or abstract-only conference records without original empirical data.
3. Studies focusing solely on bias in word embedding models, such as Word2Vec, without examining LLMs or foundation language models.
4. Studies focused solely on pretraining or fine-tuning without subsequent evaluation of model outputs, performance, or bias in a medical or health care context.
5. Studies discussing ethics, hallucination, transparency, or safety without assessing bias, fairness, or subgroup performance.
6. Studies conducted solely in medical education settings without clear clinical relevance.
7. Gray literature lacking sufficient methodological detail, clear data sources, or identifiable authorship or institutional credibility. Abstract-only records will also be excluded.

Review and Design Workflow

This review uses a staged design to enable comprehensive evidence synthesis in a rapidly expanding research field. The original search, completed in June 2025, was screened and assessed manually using predefined eligibility criteria. This phase provides a fully manually reviewed reference-standard dataset. An updated search, completed in June 2026 across the same information sources, will be used to identify additional eligible studies published since the original search.

To manage the expanded evidence base while maintaining methodological rigor, the updated search will be screened using a prespecified LLM-assisted workflow. The LLM-assisted workflow will apply the same eligibility criteria as the manual review and will be validated against the manually reviewed reference-standard dataset before interpretation of the updated search results. Studies identified as potentially eligible or uncertain by either LLM will proceed to further assessment, and final inclusion decisions will be made by at least 2 human reviewers.

This staged approach allows the review to retain the reliability of a completed manual review while also evaluating a scalable workflow for updating evidence synthesis in a fast-moving area of medical AI research. Any amendments to this protocol will be documented with a rationale and date and updated in the PROSPERO (International Prospective Register of Systematic Reviews) record.

Search Methods

This protocol follows the PRISMA-P (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols) [37] checklist to ensure methodological transparency and reproducibility. The PRISMA-P guidelines inform the study selection process, search strategy development, and reporting structure.

Confirmation of No Other Registered Reviews on This Topic

To confirm the originality of this review and prevent duplication, we searched for registered but incomplete or unpublished systematic and scoping reviews on this topic. The following databases were examined:

- CENTRAL (Cochrane Central Register of Controlled Trials)
- PROSPERO
- Evidence-Based Health Database (Epistemonikos)

No ongoing or unpublished reviews specifically addressing bias detection and mitigation in medical LLMs were identified.

Information Sources and Search Strategy

We conducted a comprehensive search of electronic databases, including Embase, MEDLINE, PsycINFO, PubMed, ACL Anthology, ACM Digital Library, arXiv, medRxiv, and bioRxiv. In addition to biomedical databases, we included ACL Anthology and ACM Digital Library to capture relevant research from computational linguistics and AI development, as LLM bias often originates from model architecture and training practices.

Search terms included a combination of MeSH and free-text keywords related to LLMs and bias in medical applications. MeSH terms were applied where available (eg, *"artificial intelligence"* AND *"Bias in health care"*), while free-text keywords (eg, *"large language model bias,"* *"GPT in medicine,"* *"AI fairness in health care"*) were used to ensure comprehensive retrieval across databases that do not support MeSH.

A broad search strategy was employed to maximize sensitivity and ensure that relevant studies were not missed. This included the use of general terms (eg, accuracy, ethics, hallucination) that may overlap with bias-related concepts. Specific inclusion criteria and screening procedures were applied to ensure that only studies directly assessing bias, fairness, or subgroup performance are included. This approach reflects the evolving and inconsistently defined nature of bias in LLM research, where relevant concepts may be described using varied terminology across disciplines.

The search strategy was refined through pilot searches and consultation with an experienced health sciences librarian from Newcastle University. Boolean operators, truncation, and wildcards were applied where appropriate to optimize retrieval. Additional search techniques included backward and forward citation searching. Reference lists of included studies were screened for additional articles, and studies citing included papers were identified using Scopus, Web of Science, and Google Scholar. Studies identified through citation searching underwent the same eligibility screening process as database search results. Full search strategies for all databases, including exact query strings, applied filters, and record counts for final searches, are provided in Multimedia Appendices 1 and 2.

The search elements are shown in Table 1.

Table 1. Search terms.

Search element	Concept	Synonyms and keywords	MeSH terms (if applicable)
Population (P)	LLMs[a]	LLM*[a] OR LMM* OR MLLM OR "Vision-Language Models" OR "Audio-Language Models" OR "Speech Language Models" OR LVLM OR VLM OR GPT*[b] OR BERT[c] OR LaMDA[d] OR PaLM[e] OR LLaMA[f] OR Claude OR Alpaca OR Falcon OR BLOOM OR "Generative AI" OR	"Natural Language Processing"[MeSH] OR "Artificial Intelligence"[MeSH]

Search element	Concept	Synonyms and keywords	MeSH terms (if applicable)
		"Transformer-based model*" OR "Foundational model*" OR OPT[g] OR Fairseq[h] OR Deepseek OR Gemini OR Med-PaLM OR Radiology-LLaMA	
Context (C)	Health care and medical applications	Health* OR Medic* OR Clinic* OR Patient* OR EHR[i] OR Physician* OR "Clinical Decision Support" OR "Medical AI" OR "Biomedical NLP[j]" OR "Biomedical Natural Language Processing" OR BioNLP[k]	"Health"[MeSH] OR "Electronic Health Records"[MeSH] OR "Clinical Decision Support Systems"[MeSH] OR "Health Services"[MeSH]
Concept (C)	Bias, fairness, ethics	Bias* OR Prejudice* OR Accuracy OR Ethic* OR Hallucination* OR Fair* OR Discrimination OR Responsible OR "Human Value Alignment" OR Inequity OR Disparit* OR Equalit* OR "Algorithmic Bias" OR "Explainability" OR "Transparency" OR "Accountability"	"Bias (Epidemiology)"[MeSH] OR "Health Equity"[MeSH] OR "Ethics, Medical"[MeSH] OR "Social Discrimination"[MeSH]

[a]LLM: large language model.
[b]GPT: Generative Pretrained Transformer.
[c]BERT: Bidirectional Encoder Representations from Transformers.
[d]LaMDA: Language Model for Dialogue Applications.
[e]PaLM: Pathways Language Model.
[f]LLaMA: Large Language Model Meta AI.
[g]OPT: Open Pretrained Transformer.
[h]FAIRSEQ: Facebook AI Research Sequence-to-Sequence Toolkit.
[i]EHR: electronic health record.
[j]NLP: natural language processing.
[k]BioNLP: biomedical natural language processing.

Study Screening and Selection

All records identified through the original June 2025 search were imported into Rayyan and deduplicated. Titles and abstracts were screened manually by members of the review team using predefined eligibility criteria. Records considered potentially eligible underwent full-text assessment. Reasons for exclusion at the full-text stage were recorded using predefined categories. Disagreements were resolved through discussion, with consultation from an additional reviewer where required.

The manually reviewed June 2025 dataset will be used as the reference standard for validating the LLM-assisted screening workflow. Records identified through the updated June 2026 search will be screened using the LLM-assisted workflow described below, followed by human verification of potentially eligible studies. The study selection process will be reported using a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram.

LLM-Assisted Screening Workflow

The LLM-assisted screening workflow is being finalized and will be specified before application to the updated search results. The final workflow will document the LLMs used, model versions, access dates, prompting strategy, input fields, classification labels, decision thresholds, and rules for resolving uncertain outputs. Two independent LLMs will screen records using the same eligibility criteria applied in the manual review. At the title and abstract stage, records will be classified as potentially eligible, excluded, or uncertain. Records classified as potentially eligible or uncertain by either LLM will progress to further assessment. Records will not be excluded on title alone.

Before interpretation of the updated search results, the LLM-assisted workflow will be evaluated against the manually reviewed June 2025 reference-standard dataset. Sensitivity will be treated as the primary screening-safety metric, because missed eligible studies represent the principal risk of an assisted screening workflow. Specificity, precision, negative predictive value, inclusion agreement, and Cohen κ will be reported as secondary metrics. Disagreements between manual and LLM-assisted decisions will be reviewed to identify potential sources of error. A sample of records excluded by the LLM-assisted workflow will undergo human audit to assess the risk of missed eligible studies, with particular attention to records using ambiguous terminology or reporting bias, fairness, or subgroup performance indirectly.

All records identified as potentially eligible through the LLM-assisted workflow will undergo final assessment by at least 2 human reviewers before inclusion. The completed systematic review will report the final LLM-assisted workflow in full, including model details, prompt templates, validation results, and any workflow modifications made before screening the updated search results.

Data Extraction

A standardized data extraction form was developed and piloted on a subset of studies to ensure clarity and consistency (Multimedia Appendix 3). Data extraction has been completed for studies included in the manual review using this form. For studies identified through the ongoing LLM-assisted workflow, data extraction will be conducted using an LLM-assisted approach based on the same predefined framework to ensure consistency across review pathways.

Any discrepancies identified during data extraction will be resolved through discussion, with involvement of an additional reviewer where necessary to achieve consensus. Extracted data will be entered into a structured spreadsheet hosted within a secure, version-controlled repository accessible to members of the review team.

Upon completion of the review, a deidentified and finalized version of the dataset will be made available through an open-access repository (eg, OSF or Zenodo) to support transparency, reproducibility, and future reuse.

Data Items

The data to be extracted are shown in Table 2.

Table 2. Data extraction categories and items.

Category	Data
Study characteristics	Title, publication date, journal, country of corresponding author, study type, source of funding, conflict of interest, primary objective
Study design	Study type (eg, experimental, observational, simulation, benchmarking), comparator(s) used (eg, human, ML[a], other LLM[b]), evaluation date
LLM model details	Model name and version (eg, GPT[c]-3.5, GPT-4 March 2023), parameter count, hyperparameters (eg, temperature, top-p), domain specificity, fine-tuning, and prompt engineering will be extracted where reported. Where such details are unavailable, they will be recorded as "not reported" and will not be used as exclusion criteria.
Training data context	Reported sources of training data (eg, PubMed, MIMIC-III[d]), whether general or specialized corpus
Medical applications	Specific medical, clinical, educational, or patient-facing tasks (eg, diagnosis, consultation, summarization)
Bias assessment methods	Bias detection methods, fairness metrics, demographic groups evaluated, impact of bias (effects on clinical reasoning, decision quality, user trust, or health care disparities)
Bias mitigation strategies	Methods used to reduce bias (eg, algorithmic debiasing, adversarial training, RAG[e]), effectiveness of mitigation
Transparency	Explainability, prompt disclosure, justification for model outputs, external validation, and transparency reporting where available
Ethical considerations	Ethical principles discussed (eg, fairness, safety, autonomy), potential harms identified, references to formal or regulatory frameworks, recommendations for responsible AI

[a]ML: machine learning.
[b]LLM: large language model.
[c]GPT: Generative Pretrained Transformer.
[d]MIMIC-III: Medical Information Mart for Intensive Care (version III).
[e]RAG: retrieval-augmented generation.

Assessment of Meta-Biases

Given the heterogeneity of study designs and outcomes anticipated in this review, quantitative assessment of publication bias (eg, funnel plots or Egger test) is unlikely to be appropriate. Instead, potential meta-biases will be assessed qualitatively. This will include comparison of findings from peer-reviewed publications and preprints, examination of selective outcome reporting (eg, incomplete reporting of subgroup or fairness analyses), and assessment of transparency in methodological reporting. The presence and potential impact of publication and reporting bias will be considered when interpreting the strength and consistency of the evidence.

Confidence in Cumulative Evidence

Formal grading of evidence using GRADE (Grading of Recommendations Assessment, Development and Evaluation) is not well suited to the objectives of this review, which focuses on mapping methodological approaches to bias detection and mitigation in medical LLMs rather than estimating intervention effects. Instead, confidence in the cumulative evidence will be assessed qualitatively based on study design, methodological transparency, robustness of bias assessment methods, consistency of findings across models and application domains, and the tailored study quality assessment framework described in this protocol. This approach aligns with guidance for evidence synthesis in emerging and methodologically heterogeneous fields.

Synthesis and Presentation of Results

The review will adhere to PRISMA guidelines, with a flow diagram detailing study selection and exclusion at each stage. Extracted data will be systematically analyzed and synthesized using a narrative approach, as the heterogeneity of study designs and methodologies precludes meta-analysis. A thematic synthesis approach will be applied, involving coding of extracted data, development of descriptive themes, and generation of analytical themes across studies. The synthesis will focus on patterns in bias detection methods, mitigation strategies, and their reported effectiveness. Ethical and governance themes identified across included studies will be incorporated into the narrative synthesis to contextualize methodological and clinical implications of bias in health care LLM applications. Studies will be grouped by health care use case (eg, clinical note summarization, diagnostic categorization), bias type (eg, demographic, clinical, systemic), bias detection methods (eg, dataset audits, counterfactual testing, adversarial evaluation), and mitigation strategies (eg, debiasing algorithms, RAG, dataset augmentation). Visual summaries, such as bar charts or heatmaps, may be used to illustrate the distribution of bias types and methodological approaches across studies.

Measures to Minimize Bias in the Review Process

The systematic review protocol includes measures to reduce selection bias, such as transparent documentation of search strategies, inclusion criteria, and data extraction methods. Any potential conflicts of interest among reviewers will be declared, and efforts will be made to ensure objectivity.

- Reviewer selection: A diverse group of reviewers with varying academic and professional backgrounds will be involved in study selection, data extraction, and synthesis to minimize bias.
- Validation framework: The LLM-assisted workflow will be evaluated against a manually reviewed reference dataset. Studies identified as eligible through the LLM-assisted workflow will undergo independent review by at least 2 human reviewers before final inclusion.
- Agreement assessment: Sensitivity will be treated as the primary screening-safety metric, as missed eligible studies represent the principal risk of an assisted screening workflow. Specificity, precision, negative predictive value, inclusion agreement, and Cohen $\varkappa$ will be reported as secondary validation metrics.
- Data interpretation: Findings will be analyzed using predefined criteria to limit subjective interpretation. Any ambiguous data will be transparently documented and resolved by consensus.
- Conflict of interest management: Reviewers will declare potential conflicts of interest before beginning. Steps will be taken to ensure impartiality, including excluding conflicted reviewers from relevant decisions if necessary.
- Transparency and reproducibility: All review processes, from search strategies to data analysis methods, will be clearly documented. Standardized tools like PRISMA will enhance rigor and transparency.
- Pilot testing: Screening, data extraction, and analysis tools will be pilot-tested to address potential biases before formal use.

Traditional clinical risk of bias tools, such as ROBINS-I [38] and QUADAS-2 [39], are not tailored for evaluating studies involving LLMs in health care. To address this gap, we developed a domain-specific framework that synthesizes principles from established tools and literature:

- AI Fairness 360 Toolkit [40]: An open-source library by IBM Research designed to detect and mitigate bias in machine learning models, providing over 70 fairness metrics and 10 bias mitigation algorithms.
- PROBAST (Prediction model Risk Of Bias Assessment Tool) [41]: A tool for assessing the risk of bias and applicability of prediction model studies, focusing on participants, predictors, outcomes, and analysis.

Our framework captures key domains relevant to LLM evaluation, including dataset transparency, bias assessment methodology, mitigation strategies, and reporting transparency. These domains represent related but distinct aspects of study quality and are considered collectively to provide a comprehensive assessment.

It was iteratively refined through discussion within the research team to ensure clarity and applicability across diverse study designs. It will also be piloted on a subset of included studies to assess consistency, with further refinements made as necessary (Table 3).

Table 3. Tailored study quality assessment framework for medical large language model studies: domains and criteria[a].

Domain	Assessment focus	Examples/operational guidance
1. Dataset transparency and representation	• Clarity on dataset sources, demographics, and limitations. Evaluates whether training and evaluation data are described and representative of clinical diversity.	• ✔️ Clearly names training/evaluation data (eg, PubMed, MIMIC-III[b]). • ✔️ Reports demographics, notes imbalances. • ✖️ No dataset source disclosed=High risk.
2. Bias assessment methods	• Whether the study actively evaluates model bias using quantitative or qualitative methods. • Studies that do not assess bias will be excluded according to eligibility criteria; therefore, this domain evaluates the robustness of bias assessment methods rather than their presence alone.	• ✔️ Subgroup analysis by race/gender/SES[c]. • ✔️ Counterfactual or adversarial evaluation. • ✖️ Limited or poorly described bias assessment methods=High risk.
3. Bias mitigation techniques	• Any method to reduce or control bias: in training, postprocessing, or model outputs.	• ✔️ Appropriate mitigation strategy applied and evaluated (eg, debiasing algorithms, adversarial reweighting, RAG[d], balanced training data) • ✖️ Mitigation strategy reported but poorly described, insufficiently justified, or not adequately evaluated=Moderate risk • ✖️ Study claims to mitigate bias but provides no clear method or evaluation, or mitigation is inadequately implemented=High risk • If mitigation is not applicable to the study aim, this domain will be coded as “not applicable” rather than contributing to overall risk.
4. Explainability, transparency, and accountability	• Does the study attempt to interpret model behavior and make it auditable?	• ✔️ Uses SHAP[e]/LIME[f] or other surrogate models. • ✔️ Applies transparency tools (eg, model cards, prompt disclosure). • ✔️ Acknowledges model uncertainty or auditability. • ✖️ Black-box outputs with no rationale=High risk.

Domain	Assessment focus	Examples/operational guidance
5. Generalizability and evaluation setting	• Contextual robustness: was the model tested across varied tasks, datasets, or populations?	• ✔️ Uses external validation datasets. • ✔️ Reports performance across demographics. • ✖️ Tested only on narrow simulation=Moderate risk.
6. Ethical, regulatory, and clinical framing	• Considers risks, benefits, alignment with AI ethics and medical governance.	• ✔️ Mentions national or international regulations or guidance. • ✔️ Discusses clinical safety, fairness, or misuse. • ✖️ No ethical or regulatory framing=High risk.

[a]Each domain will be rated as low, moderate, or high risk using predefined operational criteria. Low risk indicates comprehensive reporting and appropriate methodological implementation within the domain. Moderate risk indicates partial reporting or methodological limitations unlikely to invalidate study findings. High risk indicates substantial methodological limitations, inadequate reporting, or lack of reproducible evaluation methods likely to affect interpretability or reliability. Overall risk-of-bias judgments will be based on domain-level assessments. Studies will be classified as high risk if 2 or more domains are rated as high risk. This threshold was selected to ensure that studies with multiple methodological limitations are appropriately identified while avoiding overclassification based on a single domain. Final judgments will be confirmed through reviewer consensus. Risk levels are interpreted as follows. Low risk: fulfills low-risk standards in ≥4 domains with no domain rated high risk; Moderate risk: fulfills low-risk standards in 2 to 3 domains with no more than one domain rated high risk; High risk: two or more domains rated high risk. Studies that do not assess or report bias will be excluded according to the eligibility criteria.
[b]MIMIC-III: Medical Information Mart for Intensive Care III.
[c]SES: socioeconomic status.
[d]RAG: retrieval-augmented generation.
[e]SHAP: Shapley additive explanations.
[f]LIME: local interpretable model-agnostic explanations.

Explainability Considerations

Explainability is a critical aspect of evaluating LLMs in health care [42]. It encompasses methods that provide human-understandable insights into model decisions. While tools like Shapley additive explanations [43] and local interpretable model-agnostic explanations [44] are commonly used, they represent just one category of explainability techniques. Recent surveys highlight a broader range of approaches, including gradient- and decomposition-based methods (eg, Integrated Gradients, Layer-wise Relevance Propagation), attention visualization and probing techniques, counterfactual and data influence analyses, concept attribution methods such as testing with concept activation vectors, and natural language or chain-of-thought explanations [45,46]. Structured documentation tools like model cards and datasheets also contribute to transparency. Transparency and accountability assessment may include reporting of prompt disclosure, rationale generation, uncertainty handling, auditability, evaluator blinding, and availability of prompts or outputs where applicable. Moreover, explainability is closely linked to interpretability (understanding the internal workings of the model) and accountability (ensuring mechanisms are in place to audit and trace model decisions). Together, these practices support the responsible evaluation and integration of LLMs in clinical settings [45].

In this review, we will assess whether included studies provide any form of interpretability or transparency regarding model behavior. We will examine whether the authors used any post hoc explanation techniques, model documentation approaches, or other methods to enhance model explainability or provide rationale attribution [43,44,46]. We will also assess whether these methods are applied systematically across outputs, whether their limitations are acknowledged, and whether the study includes any audit or traceability mechanism. Studies that present model outputs without any effort to explain or justify them will be considered high risk in this domain.

Ethical Considerations

As a secondary analysis of published literature, ethical approval is not required.

Dissemination

Results will be disseminated through peer-reviewed publications, academic conferences, and open-access repositories to inform responsible LLM deployment in health care. The results of this systematic review will be published in an open-access peer-reviewed journal and shared through presentations at relevant conferences focused on medical informatics, LLMs, and health care. The findings will be made available on open-access repositories, ensuring broad accessibility. Summary reports will be created to communicate key insights to nonacademic audiences.

Patient and Public Involvement

Although patients and the public were not directly involved in the design of this systematic review, the research team has engaged with standing patient and public involvement panels at Newcastle University through other ongoing projects. These panels have previously provided input on research priorities related to AI in health care, including concerns around fairness, trust, and transparency. Insights gained from those engagements informed the broader research context, particularly considerations of bias and ethical implications in clinical applications of LLMs. We intend to disseminate a plain-language summary of the review's findings through relevant patient and public involvement networks to facilitate broader understanding and public dialogue.

Reporting Standards

This protocol was developed in accordance with the PRISMA-P 2015 statement [37], and a completed PRISMA-P checklist is included as a Checklist 1. The completed systematic review will be reported in accordance with the PRISMA 2020 guidelines [47].

Results

The phase 1 search was conducted on June 11, 2025, and yielded 15,976 records after deduplication. These records have undergone complete manual screening and assessment. The phase 2 search was conducted on June 8, 2026, and increased the total number of records to 41,143. Records from the updated search are undergoing LLM-assisted screening and eligibility assessment. Validation of the LLM-assisted workflow against the manually reviewed reference-standard dataset will be reported using sensitivity as the primary screening-safety metric, alongside specificity, precision, negative predictive value, inclusion agreement, and Cohen κ. The completed systematic review will report study selection using a PRISMA 2020 flow diagram. The review is expected to be submitted for publication in late November 2026, with subsequent publication dependent on the journal's peer-review and editorial process.

Discussion

Expected Findings

This systematic review is expected to provide a comprehensive synthesis of current approaches to bias assessment and mitigation in medical LLMs and to highlight gaps in the consistency and depth of assessment.

While existing studies often acknowledge the presence of bias, the extent to which bias is systematically evaluated and addressed remains unclear. This review will therefore examine whether bias assessment is rigorously implemented and whether mitigation strategies are actively applied and evaluated across studies.

In addition, the review will compare reported mitigation approaches and their effectiveness, with the aim of identifying patterns in current practice and areas requiring further development.

The findings are expected to inform best practices for the assessment and management of bias in medical LLMs and may contribute to the development of more structured approaches or frameworks for evaluating bias in future research, supporting the development of more transparent, equitable, and clinically reliable AI systems.

Strengths and Limitations

A key strength of this review is its comprehensive and interdisciplinary database coverage, incorporating both biomedical and computational research sources (eg, Embase, MEDLINE, PsycINFO, PubMed, ACL Anthology, ACM Digital Library, arXiv, medRxiv, and bioRxiv). The review also follows PRISMA-P guidelines and is preregistered on PROSPERO and OSF, enhancing transparency, reproducibility, and methodological rigor. In addition, the inclusion of gray literature enables the capture of emerging evidence that may not yet be available in peer-reviewed journals.

However, several limitations should be considered. Despite the inclusion of gray literature, publication bias may persist, as studies with neutral or negative findings may be underrepresented. Furthermore, heterogeneity in the definitions, measurement approaches, and reporting of bias across studies may limit comparability and preclude quantitative synthesis. Although the inclusion of preprint servers enables capture of the rapidly evolving literature on medical LLMs, studies available only as preprints have not undergone peer review and should therefore be interpreted with appropriate caution.

The LLM-assisted workflow will be validated against the completed manual review, but some risk of screening errors may remain. Relevant studies may be missed, particularly where reporting is unclear or incomplete. To reduce this risk, studies identified as eligible through the LLM-assisted workflow will undergo final review by at least 2 independent human reviewers before inclusion in the review. The completed review will report the final prompts, model versions, screening rules, validation results, and any workflow refinements made before application to the updated search results.

Acknowledgments

ChatGPT (GPT-5.6 Sol; OpenAI) was used for language editing during manuscript preparation. All scientific content and final revisions were reviewed and approved by the authors, who take full responsibility for the manuscript.

Funding

JRH is supported by an NIHR Academic Clinical Lecturership and by the Newcastle Biomedical Research Centre. EM is supported by the NIHR Newcastle Biomedical Research Centre based at the Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle University, and the Cumbria, Northumberland, and Tyne and Wear NHS Foundation Trust. AK is supported in part by the National Institute for Health and Care Research (NIHR) through an AI Award (AI_AWARD02183) and by a research grant from GlaxoSmithKline. LH is supported by the Medical Research Council (MR/T033371/1). JRH is supported by an NIHR Academic Clinical Lecturership and by the Newcastle Biomedical Research Centre. EM is supported by the NIHR Newcastle Biomedical Research Centre based at the Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle University, and the Cumbria, Northumberland, and Tyne and Wear NHS Foundation Trust. AK is supported in part by the

NIHR through an AI Award (AI_AWARD02183) and by a research grant from GlaxoSmithKline. LH is supported by the Medical Research Council (MR/T033371/1).

Data Availability

No primary data are reported in this protocol. All data generated during the review, including study selection decisions, extracted data, and quality assessments, will be made publicly available via an open-access repository (eg, OSF or Zenodo) upon publication of the completed review. Search strategies and supplementary materials will be provided as appendices to the published article.

Authors' Contributions

Conceptualization: CB, SR, JRH, AK, XYT, LH, EM, HL, HD
Methodology: JRH, EM, HL, HD
Project administration: JRH
Supervision: JRH, EM
Writing – original draft: CB, SR, JRH
Writing – review & editing: JRH, AK, XYT, LH, EM, HL, HD
All authors reviewed and approved the final manuscript.

Conflicts of Interest

JRH works for Akrivia Health as a Clinical Advisor and serves as an Expert Reviewer (Medical Devices) for the UK Medicines and Healthcare products Regulatory Agency (MHRA).
AK is supported in part by the National Institute for Health and Care Research (NIHR) through an AI Award (AI_AWARD02183) and by a research grant from GlaxoSmithKline.
EM is a guest editor for the *Artificial Intelligence in Epilepsy: Advances in Diagnosis and Treatment* collection in *Acta Epileptologica* and is the Editor-in-Chief of *JMIRx Med* (2025).
These roles and sources of support had no influence on the design or reporting of this protocol.
The other authors declare no conflicts of interest.

Multimedia Appendix 1

Full search strategy.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search results and record counts.
[DOCX File (Microsoft Word File), 18 KB-Multimedia Appendix 2]

Checklist 1

PRISMA-P 2015 checklist.
[DOCX File (Microsoft Word File), 34 KB-Checklist 1]

References

1. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI. 2018. URL: https://openai.com/index/language-unsupervised/ [Accessed 2026-01-10]
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019); Jun 2-7, 2019; Minneapolis, Minnesota, USA. URL: https://aclanthology.org/N19-1423.pdf [Accessed 2026-08-08]
3. Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. J Mach Learn Res. 2023;24:11324-11436. [doi: 10.5555/3648699.3648939]
4. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on Feb 27, 2023. [doi: 10.48550/arXiv.2302.13971]
5. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. Artif Intell Rev. 2024;57(11):299. [doi: 10.1007/s10462-024-10921-0]
6. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. iScience. May 17, 2024;27(5):109713. [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
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] [Medline: 37438534]
8. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res. Jan 7, 2025;27:e59069. [doi: 10.2196/59069] [Medline: 39773666]

9. Macia G, Liddell A, Doyle V. Conversational AI with large language models to increase the uptake of clinical guidance. Clin eHealth. Dec 2024;7:147-152. [doi: 10.1016/j.ceh.2024.12.001]
10. Vaassen B. AI, opacity, and personal autonomy. Philos Technol. Dec 2022;35(4):88. [doi: 10.1007/s13347-022-00577-5]
11. Li S, Rui H. Dual traits in probabilistic reasoning of large language models. arXiv. Preprint posted online on Dec 15, 2024. [doi: 10.48550/arXiv.2412.11009]
12. Xu H, Shuttleworth KMJ. Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". Intell Med. Feb 2024;4(1):52-57. [doi: 10.1016/j.imed.2023.08.001]
13. Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. Jun 2024;6(6):e428-e432. [doi: 10.1016/S2589-7500(24)00061-X] [Medline: 38658283]
14. Guo Y, Guo M, Su J, et al. Bias in large language models: origin, evaluation, and mitigation. Electronics (Basel). 2024;15(9):1824. [doi: 10.3390/electronics15091824]
15. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. Mar 31, 2025;43(2):1-55. [doi: 10.1145/3703155]
16. Mohottige D, Olabisi O, Boulware LE. Use of race in kidney function estimation: lessons learned and the path toward health justice. Annu Rev Med. Jan 27, 2023;74(1):385-400. [doi: 10.1146/annurev-med-042921-124419] [Medline: 36706748]
17. Tayal U, Pompei G, Wilkinson I, et al. Advancing the access to cardiovascular diagnosis and treatment among women with cardiovascular disease: a joint British Cardiovascular Societies' consensus document. Heart. Oct 28, 2024;110(22):e4. [doi: 10.1136/heartjnl-2024-324625] [Medline: 39317437]
18. Gallegos IO, Rossi RA, Barrow J, et al. Bias and fairness in large language models: a survey. In: Computational Linguistics. MIT Press; 2024:1097-1179. [doi: 10.1162/coli_a_00524]
19. Chaudhary I, Hu Q, Kumar M, Ziyadi M, Gupta R, Singh G. Certifying counterfactual bias in LLMs. Presented at: The Thirteenth International Conference on Learning Representations (ICLR 2025); Apr 24-28, 2025; Singapore. URL: https://openreview.net/forum?id=HQHnhVQznF [Accessed 2026-08-08]
20. Vats R, Agrawal S, Chippada SS. Bias detection and fairness in large language models for financial services. Int J Sci Res Comput Sci Eng Inf Technol. Mar 16, 2025;11(2):1329-1345. [doi: 10.32628/CSEIT25112461]
21. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns (N Y). Apr 12, 2024;5(4):100946. [doi: 10.1016/j.patter.2024.100946] [Medline: 38645766]
22. Hu M, Wu H, Zhu R, et al. No free lunch: retrieval-augmented generation undermines fairness in llms, even for vigilant users. In: Findings of the Association for Computational Linguistics: EMNLP 2025. Association for Computational Linguistics; 2025:18145-18170. [doi: 10.18653/v1/2025.findings-emnlp.984]
23. Zhang T, Zhou Y, Bollegala D. Evaluating the effect of retrieval augmentation on social biases. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics; 2026:5004-5026. [doi: 10.18653/v1/2026.eacl-long.233]
24. Omar M, Sorin V, Agbareia R, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health. Feb 26, 2025;24(1):57. [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
25. Kelly BS, Judge C, Bollard SM, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol. Nov 2022;32(11):7998-8007. [doi: 10.1007/s00330-022-08784-6] [Medline: 35420305]
26. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. Apr 7, 2021;4(1):65. [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]
27. Cohen JF, McInnes MDF. Deep learning algorithms to detect fractures: systematic review shows promising results but many limitations. Radiology. Jul 2022;304(1):63-64. [doi: 10.1148/radiol.212966] [Medline: 35348385]
28. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. Oct 2019;1(6):e271-e297. [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
29. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. Mar 25, 2020;368:m689. [doi: 10.1136/bmj.m689] [Medline: 32213531]
30. Vrudhula A, Kwan AC, Ouyang D, Cheng S. Machine learning and bias in medical imaging: opportunities and challenges. Circ Cardiovasc Imaging. Feb 2024;17(2):e015495. [doi: 10.1161/CIRCIMAGING.123.015495] [Medline: 38377237]

31. Koçak B, Ponsiglione A, Stanzione A, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. Diagn Interv Radiol. Mar 3, 2025;31(2):75-88. [doi: 10.4274/dir.2024.242854] [Medline: 38953330]
32. Casey A, Davidson E, Poon M, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak. Jun 3, 2021;21(1):179. [doi: 10.1186/s12911-021-01533-7] [Medline: 34082729]
33. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). Jan 21, 2025;5(1):26. [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
34. Chen SF, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine. Nat Med. Mar 2026;32(3):1152-1159. [doi: 10.1038/s41591-026-04229-5] [Medline: 41776077]
35. JBI manual for evidence synthesis. Joanna Briggs Institute; 2024. URL: https://hotus.fi/wp-content/uploads/2025/10/jbi-manual-for-evidence-synthesis-nov-2024.pdf [Accessed 2026-08-08]
36. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, California, USA. [doi: 10.5555/3295222.3295349]
37. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. Jan 1, 2015;4(1):1. [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
38. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. Oct 12, 2016;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
39. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. Oct 18, 2011;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
40. Bellamy RKE, Dey K, Hind M, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res & Dev. 2019;63(4/5):4. [doi: 10.1147/JRD.2019.2942287]
41. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. Jan 1, 2019;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
42. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9(4):e1312. [doi: 10.1002/widm.1312] [Medline: 32089788]
43. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, California, USA. [doi: 10.5555/3295222.3295230]
44. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco, California, USA. [doi: 10.1145/2939672.2939778]
45. Zhao H, Chen H, Yang F, et al. Explainability for large language models: a survey. ACM Trans Intell Syst Technol. Apr 30, 2024;15(2):1-38. [doi: 10.1145/3639372]
46. Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). Presented at: 35th International Conference on Machine Learning (ICML 2018); Jul 10-15, 2018; Stockholm, Sweden. URL: https://proceedings.mlr.press/v80/kim18d/kim18d.pdf [Accessed 2026-08-08]
47. 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] [Medline: 33782057]

Abbreviations

CENTRAL: Cochrane Central Register of Controlled Trials
Epistemonikos: Evidence-Based Health Database
GRADE: Grading of Recommendations Assessment, Development and Evaluation
LLM: large language model
PCC: Population-Concept-Context
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-P: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols
PROBAST: Prediction Model Risk of Bias Assessment Tool
PROSPERO: International Prospective Register of Systematic Reviews
RAG: retrieval-augmented generation

Edited by Javad Sarvestan; peer-reviewed by Rebecca Lin, Sivasangari Subramaniam, Yantao Xin; submitted 13.Jan.2026; final revised version received 09.Jul.2026; accepted 30.Jul.2026; published 10.Sep.2026

Please cite as:
Rao S, Boutrous C, Kormilitzin A, Tai XY, Hotchkiss L, Cong C, Meinert E, Liang H, Dong H, Harrison JR
Exploring Bias in Medical Applications of Large Language Models: Protocol for a Systematic Review
JMIR Res Protoc 2026;15:e91348
URL: https://www.researchprotocols.org/2026/1/e91348
doi: 10.2196/91348