Clinical Translation & Population Health
CLINICAL METHODS REVIEW | COMMISSIONED SYNTHESIS

Clinical Evaluation of Large-Language-Model Decision Support

Daniel Brown - Corresponding author

Abstract

This clinical methods review examines clinical evaluation of large-language-model decision support. The organizing question is what evidence is required before generated recommendations can influence diagnosis, treatment, or triage. Ten related scholarly sources are synthesized through a decision-centered framework spanning problem definition, mechanism, measurement, evaluation, implementation, and governance. The review does not invent experiments, pooled estimates, or unreported quantitative results. It instead evaluates the strength and transferability of the available evidence, with particular attention to equating vignette accuracy with safe performance in real clinical workflows. The resulting framework links technical or empirical performance to explicit use conditions and identifies tests that should precede wider adoption in clinician-facing language-model tools.

Keywords: large language models, clinical decision support, patient safety, evaluation, digital health

Synthesis lens	Principal question
Mechanism	the path from clinical data and prompt construction through model output, clinician interpretation, and patient action
Evaluation	factuality, calibration, workflow effects, subgroup safety, prospective utility, and automation bias
Primary risk	equating vignette accuracy with safe performance in real clinical workflows

1. Introduction

Research on clinical evaluation of large-language-model decision support sits at the boundary between a promising component and a consequential decision. The core question is what evidence is required before generated recommendations can influence diagnosis, treatment, or triage. Answering it requires more than assembling favorable results. It requires a chain of evidence that makes the problem boundary, mechanism, measurement, comparator, uncertainty, and accountable decision visible to the reader.

This article presents a structured narrative review of ten related publications. Sources were selected for topical relevance, traceable publication metadata, and complementary coverage of technical, empirical, and governance questions. No meta-analytic pooling is attempted because the included studies differ in designs, populations, system boundaries, and outcomes. The purpose is decision-oriented synthesis: to identify what each source can support, where transfer remains uncertain, and which evidence would justify the next stage of use.

The central risk is equating vignette accuracy with safe performance in real clinical workflows. A top-line metric or broad policy label can appear decisive while concealing the conditions that produced it. Accordingly, the synthesis separates observation from interpretation and implementation from validation. This separation is especially important when the same system creates benefits for one user or institution while transferring cost, risk, or administrative burden to another.

2. Evidence Base and Problem Boundary

A defensible review begins by defining the unit of analysis. For the present topic, the relevant boundary includes inputs, institutional or technical context, the mechanism producing an output, the person or system acting on that output, and the downstream outcome. Evidence falls outside the boundary when it measures only an intermediate proxy yet is used to claim an end-to-end benefit.

Thirunavukarasu et al. (2023) addresses "Large language models in medicine." In this synthesis, that work is used as a source on problem definition within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Meskó and Topol (2023) addresses "The imperative for regulatory oversight of large language models (or generative AI) in healthcare." In this synthesis, that work is used as a source on conceptual boundary within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Vrdoljak et al. (2025) addresses "A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration." In this synthesis, that work is used as a source on measurement within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Together, these works provide complementary entry points, but their conclusions should not be flattened into a single ranking. Differences in data generation, setting, temporal horizon, and outcome definition may be substantively meaningful. A review should therefore preserve those differences and state where analogy, rather than direct replication, connects one domain to another.

3. Mechanism and System Design

The operative mechanism is the path from clinical data and prompt construction through model output, clinician interpretation, and patient action. This formulation discourages component-only reasoning. A model, platform, policy, assay, or infrastructure service acquires practical meaning only when its interfaces and use conditions are specified. The evidence chain should describe what information is available, what transformation occurs, which assumptions make that transformation credible, and who is authorized to act.

Tripathi et al. (2024) addresses "Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care." In this synthesis, that work is used as a source on system mechanism within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Cascella et al. (2023) addresses "Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios." In this synthesis, that work is used as a source on representation choice within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Davenport and Kalakota (2019) addresses "The potential for artificial intelligence in healthcare." In this synthesis, that work is used as a source on failure analysis within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

These studies also show why modular claims require interface tests. A component may perform well while the complete pathway fails because inputs drift, users interpret outputs differently, recovery semantics are ambiguous, or institutional incentives change. Design documentation should therefore include not only the intended pathway but also foreseeable deviations, degraded modes, and conditions under which the system should stop or defer.

4. Evaluation and Uncertainty

Evaluation should center on factuality, calibration, workflow effects, subgroup safety, prospective utility, and automation bias. Average performance remains useful, but it should be accompanied by distributions, error categories, relevant subgroups or operating regimes, and uncertainty at the point where a decision changes. Comparators must isolate the value of the proposed addition rather than compare a mature intervention with an intentionally weak baseline.

Nazi and Peng (2024) addresses "Large Language Models in Healthcare and Medical Domain: A Review." In this synthesis, that

work is used as a source on comparative evaluation within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Biswas and Talukdar (2024) addresses "Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation." In this synthesis, that work is used as a source on uncertainty within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

External validation should introduce meaningful shifts in setting, time, population, workload, market structure, or environmental conditions. A nominally independent sample can still be too similar to test transfer. Conversely, a failed transfer is scientifically informative when the changed conditions are measured and the mechanism of failure is examined rather than hidden in an aggregate score.

5. Translation and Governance

Translation to clinician-facing language-model tools depends on more than technical readiness. Decision rights, documentation, maintenance, monitoring, appeal, and recovery are part of the intervention. The governance requirement is human accountability, change control, privacy protection, incident reporting, and post-deployment surveillance. Each control should be connected to a named failure mode and should identify an owner, evidence source, review interval, and escalation path.

Chen et al. (2025) addresses "Evaluating large language models and agents in healthcare: key challenges in clinical applications." In this synthesis, that work is used as a source on implementation within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Miotto et al. (2017) addresses "Deep learning for healthcare: review, opportunities and challenges." In this synthesis, that work is used as a source on governance within clinical evaluation of large-language-model decision support. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for clinician-facing language-model tools. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

A credible implementation plan also defines sunset and revalidation conditions. New data, software updates, institutional restructuring, population change, or shifts in incentives can invalidate previously reasonable assumptions. Monitoring should therefore test the validity of the evidence chain, not merely whether a service remains online or a headline metric remains stable.

6. Research Agenda

Future studies should preregister or otherwise predefine the intended decision, principal outcomes, exclusions, and analysis plan when feasible. They should report missingness, sensitivity analyses, negative or null findings, and the cost of errors to differently situated stakeholders. Reproducible artifacts should include sufficient metadata to reconstruct data provenance, model or analytical versions, parameter choices, and the sequence from evidence to conclusion.

Cross-disciplinary work needs a shared object of explanation rather than a collection of adjacent vocabularies. For clinical evaluation of large-language-model decision support, that shared object is the complete decision pathway. Technical, clinical, social, environmental, or economic expertise should each change the design or interpretation of the study. When evidence remains incomplete, the useful output is a bounded hypothesis, a transparent uncertainty statement, or a request for additional measurement.

7. Conclusion

The literature supports a disciplined approach to clinical evaluation of large-language-model decision support: define the decision, preserve system boundaries, test consequential failure modes, connect uncertainty to action, and assign accountable control. The strongest conclusion is not that one method is universally superior. It is that credible use in clinician-facing language-model tools requires an auditable link from source evidence to design choice, evaluation result, and governed decision.

Declarations

Ethics approval: Not applicable. This article synthesizes previously published literature and reports no research involving human participants, identifiable data, animals, or human tissue.

Consent to participate and consent for publication: Not applicable.

Clinical trial registration: Not applicable; no clinical intervention or participant enrollment was conducted.

Data availability: All evidence discussed in this article is identified in the reference list. No new participant-level dataset was generated.

Competing interests: The authoring group declares no competing interests for this commissioned synthesis.

References

Biswas, A., & Talukdar, W. (2024). Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation. International Journal of Innovative Science and Research Technology (IJISRT), 994-1008. https://doi.org/10.38124/ijisrt/ijisrt24may1483

Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. Journal of Medical Systems, 47(1), 33. https://doi.org/10.1007/s10916-023-01925-4

Chen, X., Xiang, J., Lu, S., Liu, Y., He, M., & Shi, D. (2025). Evaluating large language models and agents in healthcare: key challenges in clinical applications. Intelligent Medicine, 5(2), 151-163. https://doi.org/10.1016/j.imed.2025.03.002

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2), 94-98. https://doi.org/10.7861/futurehosp.6-2-94

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digital Medicine, 6(1), 120. https://doi.org/10.1038/s41746-023-00873-0

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236-1246. https://doi.org/10.1093/bib/bbx044

Nazi, Z. A., & Peng, W. (2024). Large Language Models in Healthcare and Medical Domain: A Review. Informatics, 11(3), 57. https://doi.org/10.3390/informatics11030057

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930-1940. https://doi.org/10.1038/s41591-023-02448-8

Tripathi, S., Sukumaran, R., & Cook, T. S. (2024). Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. Journal of the American Medical Informatics Association, 31(6), 1436-1440. https://doi.org/10.1093/jamia/ocad258

Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. K. (2025). A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. Healthcare, 13(6), 603. https://doi.org/10.3390/healthcare13060603