Minimum Evidence before Maternal Decision Automation

Glenn Parker, Randy Collins, Terry Griffin

Abstract

Minimum Evidence before Maternal Decision Automation addresses a narrow but consequential interpretive problem: an automated summary silently becoming a treatment decision. A bounded review of the literature is used to formulate, rather than claim to complete, an evaluation. The primary literature contributes a bounded result: the bibliographic record establishes a short correspondence item; inaccessible full-text details are not reconstructed or treated as clinical recommendations. Autonomy should be granted by task and evidence conditions, not as a global property of a model. Reversible document preparation and consequential clinical decisions carry different burdens. Define permitted operations, required review, and stop conditions. Test boundary cases with incomplete information and conflicting sources before evaluating speed. The proposed comparison concerns review triggers, incomplete records, accountable professionals; it would test whether the specific failure scenario is detected without erasing legitimate uncertainty. An integrated account should preserve the distinction between clinical correspondence and changing patient records and any stronger downstream decision.

Contrasts among the cited approaches

A useful boundary case is supplied by HIV, Pregnancy, and Infant Feeding, concerned with how evidence concerning HIV, pregnancy, and infant feeding should be interpreted in ongoing clinical communication (Xiao Yu et al. 2026c). Its methodological basis is a JAMA correspondence contribution rather than a newly reported cohort or randomized intervention. The accessible evidence is specific: the bibliographic record establishes a short correspondence item; inaccessible full-text details are not reconstructed or treated as clinical recommendations. Its relevance is the role of clinical discussion in identifying interpretation and communication obligations around sensitive longitudinal decisions. The correspondence does not provide independent new trial evidence here; this review offers no individualized HIV or feeding advice. The contribution becomes relevant to bounded autonomy when the workflow moves from review triggers to accountable professionals. That transition is the review's proposed analytical focus, not a new result attributed to the source.

A useful boundary case is supplied by Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation, concerned with how generated answers can be interpreted through the influence of their retrieved passages (Yinghao. “Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation.” *2025 4th International Conference on Robotics Sang 2025). Its methodological basis is the supplied IEEE conference title frames passage-influence analysis for explainable retrieval-augmented generation; implementation particulars were unavailable. The accessible evidence is specific: the bibliographic item is located in the official IEEE record; no unverified attribution algorithm or benchmark result is inferred. It asks for traceability between particular retrieved evidence and generated statements instead of treating a bibliography as an explanation. Passage influence, citation support, causal explanation, and factual correctness are distinct properties; the accessible record cannot establish all of them. This source enters the present review as a methodological comparison for evidence processing, not as evidence about disease biology, patient outcomes, or clinical efficacy. The transferable question is whether review triggers remains correctly interpreted after a computational choice involving accountable professionals. Its task-specific validation cannot be borrowed to validate the clinical claim.

The sources should be compared through the roles they play, not merged into a single performance claim. HIV, Pregnancy, and Infant Feeding addresses how evidence concerning HIV, pregnancy, and infant feeding should be interpreted in ongoing clinical communication (Xiao Yu et al. 2026c). By contrast, Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation addresses how generated answers can be interpreted through the influence of their retrieved passages (Yinghao. “Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation.” *2025 4th International Conference on Robotics Sang 2025). That difference matters for bounded autonomy: the evidence concerning review triggers is not automatically evidence concerning accountable professionals. A successful software operation, a plausible mechanistic account, and a validated outcome are different achievements. Their integration may improve traceability, but it cannot create an unmeasured link. A coherent review should identify where the sources complement one another and where their scopes remain incompatible. Here the proposed failure scenario is an automated summary silently becoming a treatment decision. The analytical value of that scenario is that it can occur without requiring any cited paper to be wrong: an otherwise valid local finding can be misused at a later interface. Preventing that misuse requires evidence about the interface itself, not simply a larger collection of local results.

A bounded analytical problem

Routine use of a professionally reviewed information system can turn a conditional result into an unconditional-looking recommendation. The specific concern here is bounded autonomy, not a general rejection of automation. Consider the possibility of an automated summary silently becoming a treatment decision. This is a proposed stress scenario, not an observed event reported by the cited studies. Its purpose is to expose which evidence contract would be violated even if the final output looked plausible. For clinical correspondence and changing patient records, the relevant claim must identify its population or artifact class, its time window, and the action it is supposed to support. A failure at any of these levels may persist even after a local technical improvement. The analysis therefore treats published methods as bounded evidence and asks how their limitations interact with this particular decision. An operational analogy is useful only if it keeps those limitations visible.

Autonomy should be granted by task and evidence conditions, not as a global property of a model. Reversible document preparation and consequential clinical decisions carry different burdens. Clear boundaries make useful assistance possible without confusing support with authority. In this article, the relevant contrast is review triggers, incomplete records, accountable professionals. The proposed counterexample of an automated summary silently becoming a treatment decision puts that contrast under pressure. Define permitted operations, required review, and stop conditions. Test boundary cases with incomplete information and conflicting sources before evaluating speed. A favorable finding would show that the intended interpretation of accountable professionals remains defensible when review triggers is perturbed under the specified conditions. An unfavorable finding should narrow the claim rather than be discarded as an inconvenient case.

The scenario of an automated summary silently becoming a treatment decision is designed to challenge, not illustrate success. It supplies a negative case for bounded autonomy that should be judged independently of how fluently the system explains it. The input record should retain enough information to decide which part of that scenario violates the intended contract. That includes the original source type, measurement or artifact status, and the reason the proposed interpretation exceeds its support. The evaluator would compare the system's response with a qualified judgment made without seeing that response. Disagreement should be classified rather than erased by rewriting the prompt. Clear boundaries make useful assistance possible without confusing support with authority. A useful negative result would show that the interface cannot

preserve this distinction under realistic use; a useful positive result would demonstrate preservation on independent cases rather than merely on the example used to develop the rule. The primary source remains relevant through its particular contribution: Its relevance is the role of clinical discussion in identifying interpretation and communication obligations around sensitive longitudinal decisions. That contribution should be retained even if the integrated workflow fails. Conversely, successful integration would not remove the source's limitations or turn a proposed extension into an already observed result. This asymmetric interpretation protects both the literature and the user from an unjustified all-or-nothing verdict.

Evaluation obligations

Medical machine learning must fit data provenance, care pathways, and prospective evaluation (Rajkomar, Dean, and Kohane 2019). The clinical relevance of a model depends on how its output changes a care process. Development data can contain information that becomes available only after the intended decision. A review must identify that availability boundary before comparing algorithms. Prospective use also changes who is measured and who is treated, so evaluation of a frozen historical dataset cannot establish the entire future workflow. Within maternal-infant evidence communication, this distinction is relevant to bounded autonomy. The proposed failure scenario is an automated summary silently becoming a treatment decision; its evaluation must preserve the original source's methodological scope.

Responsible health AI connects prediction with workflow, monitoring, and patient impact (Wiens et al. 2019). The point at which a model reaches practice is not the end of evaluation. New populations, changing measurements, and feedback can alter failure modes. A responsible deployment needs a route for recognizing deterioration and a person authorized to respond. Monitoring should be linked to feasible corrective actions; an alert that cannot change the workflow provides accountability only in a very limited sense. The connection to minimum evidence before maternal decision automation is a constraint on interpretation. It would be inappropriate to borrow the source's methodological authority as direct confirmation of the integrated domain claim.

Calibration separates discrimination from confidence used in clinical escalation (Guo et al. 2017). Discrimination and calibration deserve separate evaluation. A score can order cases well yet systematically overstate its certainty. Aggregate calibration can also hide local miscalibration near the threshold where action changes. An audit should retain predicted probabilities and observed outcomes rather than report only correct classifications, and it should evaluate calibration on data not reused to choose the correction. The connection to minimum evidence before maternal decision automation is a constraint on interpretation. It would be inappropriate to borrow the source's methodological authority as direct confirmation of the integrated domain claim.

BM25 remains an important lexical baseline for retrieval systems (Robertson and Zaragoza 2009). Lexical retrieval remains a useful baseline because it is explicit about matching the available terms. Semantic matching can add recall while confusing specialized senses that share broad context. A retrieval evaluation should include both exact-identifier and meaning-based cases. Comparing a new retriever with an appropriately tuned lexical baseline helps determine whether extra computation adds support rather than merely more plausible-looking passages. For the comparison involving review triggers, incomplete records, accountable professionals, the source supplies an analytical obligation. The review's proposed application remains a hypothesis, with the deliberately difficult scenario serving as a check on scope inflation.

Retrieval-augmented generation makes external evidence part of the generation process (Lewis et al. 2020). Making documents available to a generator changes its information access but does not guarantee that a statement is supported. A useful analysis separates document selection from generation conditioned on

those documents. Missing evidence, irrelevant context, and selective quotation create different failures. The proposed workflow should preserve retrieved passages and evaluate claim support independently of fluency or the mere presence of citations. This is methodological support for examining bounded autonomy, not a claim that the source studied maternal-infant evidence communication. A separate application-specific test is required at the interface described in this review.

Dense retrieval provides semantic matching but must be compared with strong lexical baselines (Karpukhin et al. 2020). Semantic retrieval can connect questions and passages whose wording differs. That advantage should be tested against cases where wording differs because the underlying population or mechanism differs. Retrieval accuracy is not identical to answer accuracy, and neither automatically establishes evidence eligibility. A good evaluation retains the ranked candidates so later errors can be localized to retrieval rather than attributed to the generator by default. For the comparison involving review triggers, incomplete records, accountable professionals, the source supplies an analytical obligation. The review's proposed application remains a hypothesis, with the deliberately difficult scenario serving as a check on scope inflation.

Replication and transparent workflows are prerequisites for reliable biomedical claims (National Academies 2019). Reproducing a calculation and independently replicating a finding are different achievements. Recoverable data and code allow one to inspect the first; independent observations and design choices inform the second. A review should avoid calling several reanalyses of one dataset separate replications. Clear reporting of analysis changes also makes a failed replication informative instead of allowing it to disappear as an implementation dispute. Within maternal-infant evidence communication, this distinction is relevant to bounded autonomy. The proposed failure scenario is an automated summary silently becoming a treatment decision; its evaluation must preserve the original source's methodological scope.

FAIR principles support reusable cohorts, signals, and analysis metadata (Wilkinson et al. 2016). Reusable evidence needs more than a downloadable file. Identifiers, metadata, provenance, and meaningful access conditions help another researcher determine what an artifact contains and whether it can answer a new question. Reuse does not mean universal permission or universal scientific comparability. The information needed to judge those boundaries should travel with the artifact rather than remain in private workflow knowledge. This is methodological support for examining bounded autonomy, not a claim that the source studied maternal-infant evidence communication. A separate application-specific test is required at the interface described in this review.

Causal inference requires a clearly specified intervention, time order, and exchangeability assumptions (Hernán and Robins 2020). A prediction can be accurate without identifying what an intervention would change. A useful causal analysis begins by defining the hypothetical comparison, eligibility, and time zero. Its assumptions concern the process generating exposure and observation, not merely the flexibility of a fitted model. In this review, causal vocabulary is appropriate only when that comparison can be specified and defended. Within maternal-infant evidence communication, this distinction is relevant to bounded autonomy. The proposed failure scenario is an automated summary silently becoming a treatment decision; its evaluation must preserve the original source's methodological scope.

Epidemiologic design makes confounding, selection, and measurement error part of interpretation (Rothman, Greenland, and Lash 2008). Confounding, selection, and measurement error are distinct problems and need not move an estimate in the same direction. An analysis should describe them before deciding which adjustment is adequate. A larger dataset can reduce random error while leaving systematic error almost untouched. Precision should therefore be interpreted alongside the quality of the design and the eligibility of its population. An implementation should expose this limitation to reviewers of clinical

correspondence and changing patient records. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

Proposed validation strategy

The question left open by the literature is whether the intended distinction survives independent checking. Define permitted operations, required review, and stop conditions. Test boundary cases with incomplete information and conflicting sources before evaluating speed. These procedures are proposed, not completed. The case set would include the scenario of an automated summary silently becoming a treatment decision, with its validity judged independently of the system's response. The relevant comparison axes are review triggers, incomplete records, accountable professionals.

A discriminating protocol would specify a preferred interpretation and at least one credible alternative before choosing measurements. It would include negative controls and observations that could favor the alternative. Reviewers would classify whether disagreement concerns data quality, model structure, or the scientific question itself. The comparison should reduce uncertainty about that distinction rather than add more observations equally compatible with every account.

The correspondence contribution should remain a source of interpretive questions, not a dataset of treatment outcomes. A source-review study can assess whether summaries preserve that role. A separate clinical or measurement study would be required to test patient outcomes, with professional oversight and appropriate ethical design. No automated-review score would substitute for that independent evidence.

Remaining uncertainty

Access limits are material for HIV, Pregnancy, and Infant Feeding, Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation. Where only bibliographic scope or a limited abstract is available, the discussion does not assign unverified datasets, algorithms, effect sizes, or clinical recommendations. This review is a bounded conceptual synthesis rather than a systematic review with an exhaustive search protocol. It cannot establish the effectiveness of the proposed integrated workflow. Its specific contribution is to articulate how bounded autonomy changes the interpretation of review triggers, incomplete records, accountable professionals. Technical sources support computational comparisons only; they do not validate biological mechanisms or patient benefit. Reviews and correspondence retain their publication roles and are not counted as new experiments. Proposed stress cases are illustrative, not observed failures attributed to the papers. Future work should therefore test the weakest inference described here, preserve contrary results, and disclose the cost of checking rather than treating additional automation as intrinsically beneficial.

Conclusion

The defensible conclusion is about the boundary between review triggers and accountable professionals, not a universal ranking of the cited methods. Define permitted operations, required review, and stop conditions. Test boundary cases with incomplete information and conflicting sources before evaluating speed. The central methodological contribution is specific: Its relevance is the role of clinical discussion in identifying interpretation and communication obligations around sensitive longitudinal decisions. Extending it to accountable professionals is an additional hypothesis, not a result supplied by the citation. A trustworthy workflow would preserve the evidence source, the assumptions of the transition, and the right to stop when support is insufficient. The proposed evaluation agenda gives those obligations observable tests while withholding claims of completed validation.

References

1. Yu, Xiao, et al. “HIV, Pregnancy, and Infant Feeding.” *JAMA*, vol. 336, no. 2, 2026, p. 169, https://doi.org/10.1001/jama.2026.6033.

2. Sang, Yinghao. “Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation.” *2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC)*, IEEE, 2025, pp. 397-400, https://doi.org/10.1109/raiic65850.2025.11170170.

3. Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. “Machine Learning in Medicine.” *New England Journal of Medicine*, vol. 380, 2019, pp. 1347-58.

4. Wiens, Jenna, et al. “Do No Harm: A Roadmap for Responsible Machine Learning for Health Care.” *Nature Medicine*, vol. 25, 2019, pp. 1337-40.

5. Guo, Chuan, et al. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1321-30.

6. Robertson, Stephen, and Hugo Zaragoza. “The Probabilistic Relevance Framework: BM25 and Beyond.” *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, 2009, pp. 333-89.

7. Lewis, Patrick, et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459-74.

8. Karpukhin, Vladimir, et al. “Dense Passage Retrieval for Open-Domain Question Answering.” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 6769-81.

9. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press, 2019.

10. Wilkinson, Mark D., et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data*, vol. 3, 2016, article 160018.

11. Hernán, Miguel A., and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

12. Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. 3rd ed., Lippincott Williams & Wilkins, 2008.