

# Model Discovery under Uneven Evidence Availability

## Abstract

Candidate metadata, accessible abstracts, absent protocols create distinct evidence obligations in exploratory hearing and age-related genetic synthesis. This literature-based conceptual analysis does not report a new experiment. The primary literature contributes a bounded result: only the supplied bibliographic scope supports discussion here; specific cohorts, variants, effect sizes, datasets, and analysis algorithms remain unverified. An unavailable method cannot be replaced by a plausible reconstruction. A title establishes subject and intention, not the implementation or result. Maintain a claim ledger separating verified details, bibliographic scope, and unanswered questions. Seek an accessible primary artifact only where a stronger claim actually requires it. The proposed comparison concerns candidate metadata, accessible abstracts, absent protocols; it would test whether the specific failure scenario is detected without erasing legitimate uncertainty. The original source's limits remain material: Without accessible methods and results, the work cannot substantiate named loci, causal direction, treatment targets, or clinical risk improvements in this review.

## Why this distinction matters

A narrow reading of model discovery under uneven evidence availability reveals a broader problem of evidence design. The question is what an informed user can conclude from candidate metadata when absent protocols remains uncertain. Consider the possibility of an inaccessible methods section filled with assumed details. 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 candidate regions and cellular annotations, 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.

An unavailable method cannot be replaced by a plausible reconstruction. A title establishes subject and intention, not the implementation or result. Evidence synthesis should make access limits visible and narrow attribution accordingly. In this article, the relevant contrast is candidate metadata, accessible abstracts, absent protocols. The proposed counterexample of an inaccessible methods section filled with assumed details puts that contrast under pressure. Maintain a claim ledger separating verified details, bibliographic scope, and unanswered questions. Seek an accessible primary artifact only where a stronger claim actually requires it. A favorable finding would show that the intended interpretation of absent protocols remains defensible when candidate metadata is perturbed under the specified conditions. An unfavorable finding should narrow the claim rather than be discarded as an inconvenient case.

A reviewer of model discovery under uneven evidence availability should be able to recover a claim before judging an output. The illustrative case is an inaccessible methods section filled with assumed details. 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. Evidence synthesis should make access limits visible and narrow attribution

accordingly. 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 stated scope separates candidate nomination and cellular plausibility from confirmed causal mechanisms. 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.

## Evidence roles and limits

(Xiao Yu et al. 2026b) examines how exploratory cross-trait analysis might nominate candidate regions shared by hearing loss and age-related disease and connect them to inner-ear contexts in Exploratory Cross-Trait Genetic Analysis of Hearing Loss and Age-Related Diseases Nominates Cardiometabolic Candidate Regions and Inner-Ear Cellular Contexts. Its methodological basis is the supplied title identifies exploratory cross-trait genetic analysis and cellular-context nomination; an authoritative abstract was not retrieved. The accessible evidence is specific: only the supplied bibliographic scope supports discussion here; specific cohorts, variants, effect sizes, datasets, and analysis algorithms remain unverified. Its stated scope separates candidate nomination and cellular plausibility from confirmed causal mechanisms. Without accessible methods and results, the work cannot substantiate named loci, causal direction, treatment targets, or clinical risk improvements in this review. For evidence access, the important connection is the distinction between candidate metadata and absent protocols. The source informs that distinction, but a stronger conclusion would require evidence at the point where one is used to stand for the other.

The relevant question in BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models is how reinforcement learning for program repair can overcome sparse execution feedback and coarse sequence-level credit (Yuanhao Li et al. 2026a). Its methodological basis is a three-stage pipeline combining execution-verified supervised traces, sequence- and line-level reward models, and PPO with reward redistribution to critical edit regions. The accessible evidence is specific: evaluation spans SWE-bench Verified, Defects4J transfer, HumanEval-Java, and QuixBugs, with reported gains in both repository repair and cross-language generalization. Line-level credit assignment makes test execution informative at the granularity where patches actually change behavior. Benchmark success does not by itself establish maintainability, specification fidelity, or robustness to incomplete and flaky tests. 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 candidate metadata remains correctly interpreted after a computational choice involving absent protocols. Its task-specific validation cannot be borrowed to validate the clinical claim.

The relevant question in Evo-CuRL: Curriculum-Aware Reinforcement Learning over Code Lineage Graphs for Software Engineering Reasoning is how software-engineering reasoning can learn from the evolving ancestry of code states rather than isolated prompts and patches (Wei Tan et al. 2026). Its methodological basis is curriculum-aware reinforcement learning organized over code-lineage graphs, so training can expose increasingly difficult relationships among revisions and outcomes. The accessible evidence is specific: the conference record establishes a graph-and-curriculum formulation for software reasoning, with evaluation reported in the software-engineering setting. The lineage representation offers a natural way to retain failed branches, successful repairs, and prerequisite relations during learning. The value of a curriculum depends on graph construction, difficulty ordering, and whether benchmark lineages resemble real collaborative repositories. 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 candidate metadata remains correctly interpreted after a computational choice involving absent protocols. Its task-specific validation cannot be borrowed to validate the clinical claim.

A useful boundary case is supplied by *Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs*, concerned with how fine-tuned language models remain reliable when retrieved context is noisy or irrelevant (Yinghao. “Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs.” \*2025 6th International Conference on Artificial Intelligence Sang 2025). Its methodological basis is the supplied IEEE conference title frames retrieval-noise robustness in fine-tuned LLMs; detailed methods were not accessible. The accessible evidence is specific: the official IEEE document record confirms the bibliographic item but does not establish a verified noise protocol or quantitative result in the accessible material. Its stated problem makes the retrieval interface, rather than only model weights, a locus of robustness evaluation. No specific training method, benchmark, noise rate, or effect size is attributed without the full technical record. 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 candidate metadata remains correctly interpreted after a computational choice involving absent protocols. 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. *Exploratory Cross-Trait Genetic Analysis of Hearing Loss and Age-Related Diseases Nominates Cardiometabolic Candidate Regions and Inner-Ear Cellular Contexts* addresses how exploratory cross-trait analysis might nominate candidate regions shared by hearing loss and age-related disease and connect them to inner-ear contexts (Xiao Yu et al. 2026b). By contrast, *Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs* addresses how fine-tuned language models remain reliable when retrieved context is noisy or irrelevant (Yinghao. “Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs.” \*2025 6th International Conference on Artificial Intelligence Sang 2025). That difference matters for evidence access: the evidence concerning candidate metadata is not automatically evidence concerning absent protocols. 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 inaccessible methods section filled with assumed details. 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.

## Testing the interpretation

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. An implementation should expose this limitation to reviewers of candidate regions and cellular annotations. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

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. An implementation should expose this limitation to reviewers of candidate regions and cellular annotations. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

Policy distillation connects compression with sequential decision behavior (Rusu et al. 2015). Compressing a policy concerns sequential behavior rather than one isolated prediction. A small local mismatch can compound across actions, while a locally similar policy can explore differently. Evaluation should retain trajectories and consequential state transitions. The review treats policy transfer as a potential operational tool whose task-level contract still needs checking in the destination environment. For the comparison involving candidate metadata, accessible abstracts, absent protocols, 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.

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. Within exploratory hearing and age-related genetic synthesis, this distinction is relevant to evidence access. The proposed failure scenario is an inaccessible methods section filled with assumed details; its evaluation must preserve the original source's methodological scope.

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 evidence access, not a claim that the source studied exploratory hearing and age-related genetic synthesis. 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. The connection to model discovery under uneven evidence availability is a constraint on interpretation. It would be inappropriate to borrow the source's methodological authority as direct confirmation of the integrated domain claim.

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 exploratory hearing and age-related genetic

synthesis, this distinction is relevant to evidence access. The proposed failure scenario is an inaccessible methods section filled with assumed details; 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. An implementation should expose this limitation to reviewers of candidate regions and cellular annotations. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

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 exploratory hearing and age-related genetic synthesis, this distinction is relevant to evidence access. The proposed failure scenario is an inaccessible methods section filled with assumed details; 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 candidate regions and cellular annotations. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

## **Proposed comparisons and controls**

The proposed test of evidence access is concrete: Maintain a claim ledger separating verified details, bibliographic scope, and unanswered questions. Seek an accessible primary artifact only where a stronger claim actually requires it. These procedures are proposed, not completed. The case set would include the scenario of an inaccessible methods section filled with assumed details, with its validity judged independently of the system's response. The relevant comparison axes are candidate metadata, accessible abstracts, absent protocols.

An evidence-role protocol would map each publication to its underlying observations and its permissible conclusions. The test cases would include reviews of the same primary study, correspondence without new data, and observations at incompatible scales. Reviewers would identify scope inflation before seeing the automated summary. An integrated system would be credited for preserving a missing link, not rewarded for making every source fit one complete narrative.

For the exploratory genetic source, obtaining and inspecting the primary methods is a prerequisite to stronger attribution. The proposed review can presently test evidence-role preservation, but it cannot validate unverified genetic regions, sample construction, or tissue mappings. Any later biological protocol would

need to be designed around the actual reported candidates rather than invented examples assigned to the publication.

### **Limits of the synthesis**

Access limits are material for Exploratory Cross-Trait Genetic Analysis of Hearing Loss and Age-Related Diseases Nominates Cardiometabolic Candidate Regions and Inner-Ear Cellular Contexts, Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs. 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 evidence access changes the interpretation of candidate metadata, accessible abstracts, absent protocols. 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

Candidate metadata deserves to remain distinguishable from absent protocols. That distinction is the central safeguard identified by this review. Maintain a claim ledger separating verified details, bibliographic scope, and unanswered questions. Seek an accessible primary artifact only where a stronger claim actually requires it. The remaining problem can be stated as a testable contrast involving candidate metadata, accessible abstracts, absent protocols, with an inaccessible methods section filled with assumed details as a deliberately challenging alternative. Further work should test the comparison on independent cases, measure correction and review costs, and retain failed transitions as evidence. Until then, the integrated account is a research agenda rather than a new empirical guarantee.

## References

1. Yu, Xiao, et al. “Exploratory Cross-Trait Genetic Analysis of Hearing Loss and Age-Related Diseases Nominates Cardiometabolic Candidate Regions and Inner-Ear Cellular Contexts.” *Intelligent Translational Medicine*, 2026, <https://doi.org/10.1016/j.itm.2026.100002>.
2. Li, Yuanhao, et al. “BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models.” *arXiv*, version 4, 18 Sept. 2026, <https://arxiv.org/abs/2605.09134v4>. Preprint.
3. Tan, Wei, et al. “Evo-CuRL: Curriculum-Aware Reinforcement Learning over Code Lineage Graphs for Software Engineering Reasoning.” *Proceedings of the 2026 International Conference on Multimedia Retrieval*, ACM, 2026, pp. 1327-35, <https://doi.org/10.1145/3805622.3810701>.
4. Sang, Yinghao. “Robustness of Fine-Tuned LLMs under Noisy Retrieval Inputs.” *2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, IEEE, 2025, pp. 417-20, <https://doi.org/10.1109/aiea66061.2025.11160531>.
5. Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. “Machine Learning in Medicine.” *New England Journal of Medicine*, vol. 380, 2019, pp. 1347-58.
6. Wiens, Jenna, et al. “Do No Harm: A Roadmap for Responsible Machine Learning for Health Care.” *Nature Medicine*, vol. 25, 2019, pp. 1337-40.
7. Rusu, Andrei A., et al. “Policy Distillation.” *arXiv*, 2015, arXiv:1511.06295.
8. 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.
9. Lewis, Patrick, et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459-74.
10. 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.
11. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press, 2019.
12. Wilkinson, Mark D., et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data*, vol. 3, 2016, article 160018.
13. Hernán, Miguel A., and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.
14. Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. 3rd ed., Lippincott Williams & Wilkins, 2008.