Repairing Audiometry Pipelines without Rewriting Clinical Meaning

Evan Wallace, Noah Spencer, Liam Gardner

Abstract

Test fixtures, frequency units, clinical labels create distinct evidence obligations in occupational hearing-risk assessment. This literature-based conceptual analysis does not report a new experiment. The primary literature contributes a bounded result: the accessible publication describes 7139 initially recruited noise-exposed subjects and longitudinal and external evaluation; baseline 3-kHz and 3-6-kHz measures contribute prognostic information. A computational specification is a translation of a substantive question. Errors can occur in that translation even when every implementation step is correct. Build examples with known substantive answers, including boundary cases. Review diffs against a written endpoint dictionary and require approval for meaning-changing patches. The proposed comparison concerns test fixtures, frequency units, clinical labels; it would test whether the specific failure scenario is detected without erasing legitimate uncertainty. The original source's limits remain material: Limited follow-up, a smaller longitudinal sample, occupational selection, and untested prospective clinical impact constrain deployment claims.

Source-specific contributions

Prediction of Risk of Hearing Loss by Industry Noise from Cross-Sectional and Longitudinal Data asks whether baseline hearing measurements can predict subsequent speech-frequency hearing loss in noise-exposed industrial workers (Xiao Yu et al. 2026d). Its methodological basis is sex-stratified cross-sectional modeling, longitudinal generalized estimating equations, and Cox-model nomograms using baseline high-frequency audiometry. The accessible evidence is specific: the accessible publication describes 7139 initially recruited noise-exposed subjects and longitudinal and external evaluation; baseline 3-kHz and 3-6-kHz measures contribute prognostic information. The design distinguishes association with existing impairment from prediction of later impairment. Limited follow-up, a smaller longitudinal sample, occupational selection, and untested prospective clinical impact constrain deployment claims. The source supplies one part of the comparison involving test fixtures, frequency units, clinical labels. Its evidence should retain its original unit and scope; the present synthesis proposes a connection without claiming that the paper directly tested this integrated workflow.

A useful boundary case is supplied by BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models, concerned with 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 test fixtures remains correctly interpreted after a computational choice involving clinical labels. Its task-specific validation cannot be borrowed to validate the clinical claim.

(Wei Tan et al. 2026) examines how software-engineering reasoning can learn from the evolving ancestry of code states rather than isolated prompts and patches in Evo-CuRL: Curriculum-Aware Reinforcement Learning over Code Lineage Graphs for Software Engineering Reasoning. 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 test fixtures remains correctly interpreted after a computational choice involving clinical labels. Its task-specific validation cannot be borrowed to validate the clinical claim.

The relevant question in Adaptive Quantization Strategies for Robust ML Inference under Distribution Shift is how an inference system should revise its quantization policy when deployment data no longer resembles calibration data (Yinghao. “Adaptive Quantization Strategies for Robust ML Inference under Distribution Shift.” *Proceedings of the 2026 5th International Conference on Cyber Security Sang 2026). Its methodological basis is an adaptive quantization strategy that treats numerical precision as a deployment control rather than a fixed compression choice. The accessible evidence is specific: the study is framed around robustness under distribution shift and therefore makes the stability of low-precision inference, not compression alone, the central outcome. It connects efficient inference with the broader problem of shift-aware reliability. Its conclusions should be tested across architectures, bit widths, hardware kernels, and shifts that were not represented during calibration. 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 test fixtures remains correctly interpreted after a computational choice involving clinical labels. 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. Prediction of Risk of Hearing Loss by Industry Noise from Cross-Sectional and Longitudinal Data addresses whether baseline hearing measurements can predict subsequent speech-frequency hearing loss in noise-exposed industrial workers (Xiao Yu et al. 2026d). By contrast, Adaptive Quantization Strategies for Robust ML Inference under Distribution Shift addresses how an inference system should revise its quantization policy when deployment data no longer resembles calibration data (Yinghao. “Adaptive Quantization Strategies for Robust ML Inference under Distribution Shift.” *Proceedings of the 2026 5th International Conference on Cyber Security Sang 2026). That difference matters for specification fidelity: the evidence concerning test fixtures is not automatically evidence concerning clinical labels. 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 a patch passing tests while converting units incorrectly. 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.

Framing the scientific question

The most informative comparison in occupational hearing-risk assessment may occur before a score is calculated. Decisions about test fixtures, frequency units, and clinical labels already determine what the score will mean. Consider the possibility of a patch passing tests while converting units incorrectly. 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 noise-exposed workers and clinical follow-up, 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.

A computational specification is a translation of a substantive question. Errors can occur in that translation even when every implementation step is correct. Repairing code requires preserving the intended denominator, units, exclusions, and endpoint logic. In this article, the relevant contrast is test fixtures, frequency units, clinical labels. The proposed counterexample of a patch passing tests while converting units incorrectly puts that contrast under pressure. Build examples with known substantive answers, including boundary cases. Review diffs against a written endpoint dictionary and require approval for meaning-changing patches. A favorable finding would show that the intended interpretation of clinical labels remains defensible when test fixtures is perturbed under the specified conditions. An unfavorable finding should narrow the claim rather than be discarded as an inconvenient case.

The strongest comparison would isolate the disputed transition in test fixtures, frequency units, clinical labels. The proposed counterexample, a patch passing tests while converting units incorrectly, identifies what should not be accepted as equivalent evidence. 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. Repairing code requires preserving the intended denominator, units, exclusions, and endpoint logic. 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: The design distinguishes association with existing impairment from prediction of later impairment. 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.

From assumptions to observable checks

False-discovery control is central when many molecular features are screened (Benjamini and Hochberg 1995). Multiplicity arises from the family of questions actually searched. Controlling false discoveries within a named family does not automatically cover later tissue filters, alternative endpoints, or narrative choices. The proposed audit should retain exploratory branches so the family is visible. Independent confirmation then addresses a different question from whether a discovery threshold was correctly applied during screening. An implementation should expose this limitation to reviewers of noise-exposed workers and clinical follow-up. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

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. For the comparison involving test fixtures, frequency units, clinical labels, 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.

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 repairing audiometry pipelines without rewriting clinical meaning is a constraint on interpretation. It would be inappropriate to borrow the source's methodological authority as direct confirmation of the integrated domain claim.

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. An implementation should expose this limitation to reviewers of noise-exposed workers and clinical follow-up. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

Proximal policy optimization offers a widely used objective for stable policy updates (Schulman et al. 2017). An optimization procedure can make policy updates tractable without deciding whether the reward represents the desired behavior. Stable training and legitimate task success are separate properties. A review should inspect reward boundaries, information leakage, and held-out failure modes. A larger rewarded score is weak evidence if the policy can improve it by selecting easier tasks or hiding unresolved outputs. The connection to repairing audiometry pipelines without rewriting clinical meaning is a constraint on interpretation. It would be inappropriate to borrow the source's methodological authority as direct confirmation of the integrated domain claim.

The information bottleneck formalizes compression that retains task-relevant information (Tishby, Pereira, and Bialek 1999). Compression is useful when it retains the information needed for the intended task. That task must be specified before one can judge what information is expendable. A bottleneck can discard nuisance variation but can also discard rare distinctions important to another user. The analysis should compare retained task information with the explicit downstream requirement rather than treating brevity as self-evidently beneficial. The connection to repairing audiometry pipelines without rewriting clinical meaning is a constraint on interpretation. It would be inappropriate to borrow the source's methodological authority as direct confirmation of the integrated domain claim.

Variational bottlenecks make relevance-compression tradeoffs trainable in deep networks (Alemi et al. 2017). A trainable relevance-compression tradeoff needs an operational account of relevance. If labels omit a consequential distinction, a model can satisfy the training objective by discarding it. Hyperparameter comparisons should therefore include boundary and rare cases. The review proposes interpreting a

variational bottleneck as a controlled modeling choice whose benefits and losses must be measured separately. For the comparison involving test fixtures, frequency units, clinical labels, 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. For the comparison involving test fixtures, frequency units, clinical labels, 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.

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. Within occupational hearing-risk assessment, this distinction is relevant to specification fidelity. The proposed failure scenario is a patch passing tests while converting units incorrectly; its evaluation must preserve the original source's methodological scope.

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. This is methodological support for examining specification fidelity, not a claim that the source studied occupational hearing-risk assessment. A separate application-specific test is required at the interface described in this review.

An agenda for independent evaluation

The proposed test of specification fidelity is concrete: Build examples with known substantive answers, including boundary cases. Review diffs against a written endpoint dictionary and require approval for meaning-changing patches. These procedures are proposed, not completed. The case set would include the scenario of a patch passing tests while converting units incorrectly, with its validity judged independently of the system's response. The relevant comparison axes are test fixtures, frequency units, clinical labels.

A contract-focused protocol would pair executable fixtures with written substantive requirements. Fixtures would cover units, exclusions, provenance, interruptions, and boundary cases. The evaluation would preserve the first interface at which a requirement fails and separate that failure from its downstream symptoms. A repaired or successfully replayed computation would not receive credit for scientific validity unless the requirement itself still expresses the intended question.

An occupational study would need decision-time eligibility, subsequent outcome observation, and a realistic follow-up pathway. Model comparison should account for worker turnover and baseline impairment. Calibration in a validation cohort is informative, but improved prevention or referral quality would require evaluation of the action policy and its burdens, not a higher discrimination statistic alone.

Transfer and access limitations

Access limits are material for Adaptive Quantization Strategies for Robust ML Inference under Distribution Shift. 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 specification fidelity changes the interpretation of test fixtures, frequency units, clinical labels. 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

An analysis of repairing audiometry pipelines without rewriting clinical meaning should end with a bounded claim: useful evidence can be preserved without pretending that every downstream inference has been tested. Build examples with known substantive answers, including boundary cases. Review diffs against a written endpoint dictionary and require approval for meaning-changing patches. The primary source remains bounded by its own design: Limited follow-up, a smaller longitudinal sample, occupational selection, and untested prospective clinical impact constrain deployment claims. 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. “Prediction of Risk of Hearing Loss by Industry Noise from Cross-Sectional and Longitudinal Data.” *Communications Medicine*, vol. 6, 2026, https://doi.org/10.1038/s43856-026-01463-3.

2. Li, Y., Wang, H., Shang, X., Tang, X., Cao, Y., & Chen, X. (2026). BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models. arXiv preprint arXiv:2605.09134.

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. “Adaptive Quantization Strategies for Robust ML Inference under Distribution Shift.” *Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy*, ACM, 2026, pp. 362-68, https://doi.org/10.1145/3819836.3819892.

5. Benjamini, Yoav, and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B*, vol. 57, no. 1, 1995, pp. 289-300.

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

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

8. Rusu, Andrei A., et al. “Policy Distillation.” *arXiv*, 2015, arXiv:1511.06295.

9. Schulman, John, et al. “Proximal Policy Optimization Algorithms.” *arXiv*, 2017, arXiv:1707.06347.

10. Tishby, Naftali, Fernando C. Pereira, and William Bialek. “The Information Bottleneck Method.” *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 1999, pp. 368-77.

11. Alemi, Alexander A., et al. “Deep Variational Information Bottleneck.” *International Conference on Learning Representations*, 2017.

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

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

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