Oxidative Stress beside Prognosis: Avoiding Endpoint Substitution

Jack Peterson, Luke Richardson, Adam Jenkins

Abstract

This conceptual analysis examines endpoint alignment in occupational hearing-risk assessment, using stroke markers, hearing forecasts, systemic risk as the key comparison axes. This literature-based conceptual analysis does not report a new experiment. The starting source uses sex-stratified cross-sectional modeling, longitudinal generalized estimating equations, and Cox-model nomograms using baseline high-frequency audiometry. An endpoint deserves the interpretation attached to its measurement procedure and time horizon. A workflow completion, biomarker response, or early recovery score cannot replace a patient-level benefit without additional evidence. Write an endpoint dictionary with units, timing, population, and decision role. Examine substitutions explicitly and ask which conclusions would survive if the proxy were imperfect. The proposed comparison concerns stroke markers, hearing forecasts, systemic risk; it would test whether the specific failure scenario is detected without erasing legitimate uncertainty. An integrated account should preserve the distinction between noise-exposed workers and clinical follow-up and any stronger downstream decision.

A bounded analytical problem

The most informative comparison in occupational hearing-risk assessment may occur before a score is calculated. Decisions about stroke markers, hearing forecasts, and systemic risk already determine what the score will mean. Consider the possibility of a vascular biomarker standing in for hearing-loss prevention. 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.

An endpoint deserves the interpretation attached to its measurement procedure and time horizon. A workflow completion, biomarker response, or early recovery score cannot replace a patient-level benefit without additional evidence. Alignment connects the measured event to the question being answered. In this article, the relevant contrast is stroke markers, hearing forecasts, systemic risk. The proposed counterexample of a vascular biomarker standing in for hearing-loss prevention puts that contrast under pressure. Write an endpoint dictionary with units, timing, population, and decision role. Examine substitutions explicitly and ask which conclusions would survive if the proxy were imperfect. A favorable finding would show that the intended interpretation of systemic risk remains defensible when stroke markers is perturbed under the specified conditions. An unfavorable finding should narrow the claim rather than be discarded as an inconvenient case.

A reviewer of oxidative stress beside prognosis: avoiding endpoint substitution should be able to recover a claim before judging an output. The illustrative case is a vascular biomarker standing in for hearing-loss prevention. 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. Alignment connects the measured event to the question being answered. 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.

Evaluation obligations

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 stroke markers, hearing forecasts, systemic risk, 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 stroke markers, hearing forecasts, systemic risk, 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. This is methodological support for examining endpoint alignment, 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.

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. The connection to oxidative stress beside prognosis: avoiding endpoint substitution is a constraint on interpretation. It would be inappropriate to borrow the source's methodological authority as direct confirmation of the integrated domain claim.

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. This is methodological support for examining endpoint alignment, 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.

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. For the comparison involving stroke markers, hearing forecasts, systemic risk, 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.

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 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.

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. Within occupational hearing-risk assessment, this distinction is relevant to endpoint alignment. The proposed failure scenario is a vascular biomarker standing in for hearing-loss prevention; its evaluation must preserve the original source's methodological scope.

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. Within occupational hearing-risk assessment, this distinction is relevant to endpoint alignment. The proposed failure scenario is a vascular biomarker standing in for hearing-loss prevention; its evaluation must preserve the original source's methodological scope.

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. This is methodological support for examining endpoint alignment, 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.

Contrasts among the cited approaches

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. For endpoint alignment, the important connection is the distinction between stroke markers and systemic risk. The source informs that distinction, but a stronger conclusion would require evidence at the point where one is used to stand for the other.

A useful boundary case is supplied by A New Playing Method of the Guessing Football Lottery, concerned with how buyer opinions and alternative ticket structures might affect football-lottery prediction and betting behavior (Shunqi Liu et al. 2020). Its methodological basis is a proposed 'sky ladder' playing strategy evaluated through feasibility analysis and simulations using back-propagation and LSTM neural networks. The accessible evidence is specific: the reported simulations are used to argue that the new format offers flexibility while controlling the risks faced by buyers and lottery operators. The paper links predictive modeling with product design and behavioral response instead of treating match probabilities as the only object of study. Simulation performance cannot establish bettor welfare or market effects without prospective calibration, behavioral data, and comparison against simple probabilistic baselines. 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 stroke markers remains correctly interpreted after a computational choice involving systemic risk. Its task-specific validation cannot be borrowed to validate the clinical claim.

(Yuanhao Li et al. 2026a) examines how reinforcement learning for program repair can overcome sparse execution feedback and coarse sequence-level credit in BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models. 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 stroke markers remains correctly interpreted after a computational choice involving systemic risk. Its task-specific validation cannot be borrowed to validate the clinical claim.

(CongCong Wu et al. 2026) examines how oxidative-stress-related evidence can link noise exposure and stroke across human and experimental settings in An Oxidative Stress Biomarker for Noise-Associated Stroke: Evidence from Human and Mice. Its methodological basis is gene-expression screening with machine-learning approaches, longitudinal human analysis, a four-week noise-exposed mouse experiment, and NHANES observational analysis. The accessible evidence is specific: the abstract identifies CCL3 as a candidate and reports experimental gene checks and associations of oxidative-balance scores in noise-exposed participants. The paper combines several evidence scales rather than relying on one observational correlation. Triangulation does not eliminate confounding or species differences, and oxidative-balance associations are not evidence that a particular dietary intervention prevents stroke. Reading the work through endpoint alignment changes the evaluation question: the issue is whether stroke markers supports the intended interpretation of systemic risk, rather than whether the reported method is impressive in its original task.

(Yinghao. “Adaptive Quantization Strategies for Robust ML Inference under Distribution Shift.” *Proceedings of the 2026 5th International Conference on Cyber Security Sang 2026) examines how an inference system should revise its quantization policy when deployment data no longer resembles calibration data in Adaptive Quantization Strategies for Robust ML Inference under Distribution Shift. 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 stroke markers remains correctly interpreted after a computational choice involving systemic risk. 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 endpoint alignment: the evidence concerning stroke markers is not automatically evidence concerning systemic risk. 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 vascular biomarker standing in for hearing-loss prevention. 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.

Proposed validation strategy

Write an endpoint dictionary with units, timing, population, and decision role. Examine substitutions explicitly and ask which conclusions would survive if the proxy were imperfect. This would operationalize the concern with endpoint alignment rather than assume it has been resolved by one successful output. These procedures are proposed, not completed. The case set would include the scenario of a vascular biomarker standing in for hearing-loss prevention, with its validity judged independently of the system's response. The relevant comparison axes are stroke markers, hearing forecasts, systemic risk.

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.

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.

Remaining uncertainty

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 endpoint alignment changes the interpretation of stroke markers, hearing forecasts, systemic risk. 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 stroke markers and systemic risk, not a universal ranking of the cited methods. Write an endpoint dictionary with units, timing, population, and decision role. Examine substitutions explicitly and ask which conclusions would survive if the proxy were imperfect. The central methodological contribution is specific: The design distinguishes association with existing impairment from prediction of later impairment. Extending it to systemic risk is an additional hypothesis, not a result supplied by the citation. 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. “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. Liu, Shunqi, et al. “A New Playing Method of the Guessing Football Lottery.” *IOP Conference Series: Materials Science and Engineering*, vol. 790, no. 1, 2020, https://doi.org/10.1088/1757-899x/790/1/012100.

3. 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.

4. Wu, CongCong, et al. “An Oxidative Stress Biomarker for Noise-Associated Stroke: Evidence from Human and Mice.” *Noise and Health*, vol. 28, no. 133, 2026, pp. 945-58, https://doi.org/10.4103/nah.nah_239_25.

5. 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.

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

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

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

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

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

11. 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.

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

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

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

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