Counting Independent Evidence in Mechanistic Reviews

Victor Wheeler, Ernest Spencer, Harold Davidson

Abstract

Shared cohorts, overlapping reviews, source lineage create distinct evidence obligations in immune-state and tumor evidence synthesis. A bounded review of the literature is used to formulate, rather than claim to complete, an evaluation. The starting source uses a conceptual review proposing immunometabolic set points across immune and tissue compartments. Sources often reuse cohorts, measurements, and earlier interpretations. Counting publications can overstate the number of independent tests. Map datasets and overlapping samples before synthesis. Weight replication by independent data and design differences rather than by bibliography length. The proposed comparison concerns shared cohorts, overlapping reviews, source lineage; it would test whether the specific failure scenario is detected without erasing legitimate uncertainty. The original source's limits remain material: A unifying set point is a hypothesis, not an already validated clinical measurement or an established therapeutic protocol.

The failure scenario

The most informative comparison in immune-state and tumor evidence synthesis may occur before a score is calculated. Decisions about shared cohorts, overlapping reviews, and source lineage already determine what the score will mean. Consider the possibility of several summaries counted as independent experiments. 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 tissue contexts, clinical endpoints, and mechanistic claims, 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.

Sources often reuse cohorts, measurements, and earlier interpretations. Counting publications can overstate the number of independent tests. Evidence independence should be assessed at the underlying observation and hypothesis level. In this article, the relevant contrast is shared cohorts, overlapping reviews, source lineage. The proposed counterexample of several summaries counted as independent experiments puts that contrast under pressure. Map datasets and overlapping samples before synthesis. Weight replication by independent data and design differences rather than by bibliography length. A favorable finding would show that the intended interpretation of source lineage remains defensible when shared cohorts is perturbed under the specified conditions. An unfavorable finding should narrow the claim rather than be discarded as an inconvenient case.

The scenario of several summaries counted as independent experiments is designed to challenge, not illustrate success. It supplies a negative case for evidence independence 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. Evidence independence should be assessed at the underlying observation and hypothesis level. 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 set-point framing asks whether durable remission requires changing a coupled metabolic state rather than only lowering an inflammatory marker. 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.

Reading the literature by evidence type

The relevant question in Resetting Immunometabolic Set Points in Autoimmune Disease is why autoimmune activity can persist or recur when conventional inflammation suppression does not durably change underlying cellular and tissue states (Xiao Yu et al. 2026e). Its methodological basis is a conceptual review proposing immunometabolic set points across immune and tissue compartments. The accessible evidence is specific: the review integrates cellular metabolism, nutrient/oxygen conditions, mitochondrial stress, and stromal metabolites as candidate mechanisms of persistence. The set-point framing asks whether durable remission requires changing a coupled metabolic state rather than only lowering an inflammatory marker. A unifying set point is a hypothesis, not an already validated clinical measurement or an established therapeutic protocol. For evidence independence, the important connection is the distinction between shared cohorts and source lineage. The source informs that distinction, but a stronger conclusion would require evidence at the point where one is used to stand for the other.

(Sunjiang Liu et al. 2026) examines whether accessible preoperative biomarkers can predict early visual-field nonrecovery after surgery for nonfunctioning pituitary adenomas in Machine Learning-Based Prediction of Early Visual Field Nonrecovery after Endoscopic Surgery for Nonfunctioning Pituitary Adenomas with Clinically Accessible Biomarkers: A Dual-Center Study. Its methodological basis is a retrospective dual-center design with LASSO selection, multivariable logistic modeling, and a nomogram; training comprised 268 patients and validation 100. The accessible evidence is specific: the study includes 368 patients, defines early improvement at postoperative day seven, and reports age, vertical tumor diameter, and fasting glucose as predictors; reported C-indices are approximately 0.747 and 0.753. It develops an accessible early-risk tool with geographically separate validation instead of relying solely on model fit in development data. Day-seven nonrecovery is not permanent visual loss; retrospective selection, limited validation events, and unmeasured care differences constrain interpretation. The source supplies one part of the comparison involving shared cohorts, overlapping reviews, source lineage. 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.

The sources should be compared through the roles they play, not merged into a single performance claim. Resetting Immunometabolic Set Points in Autoimmune Disease addresses why autoimmune activity can persist or recur when conventional inflammation suppression does not durably change underlying cellular and tissue states (Xiao Yu et al. 2026e). By contrast, Machine Learning-Based Prediction of Early Visual Field Nonrecovery after Endoscopic Surgery for Nonfunctioning Pituitary Adenomas with Clinically Accessible Biomarkers: A Dual-Center Study addresses whether accessible preoperative biomarkers can predict early visual-field nonrecovery after surgery for nonfunctioning pituitary adenomas (Sunjiang Liu et al. 2026). That difference matters for evidence independence: the evidence concerning shared cohorts is not automatically evidence concerning source lineage. 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 several summaries

counted as independent experiments. 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.

Methods that constrain the claim

Uncertainty can deteriorate under the site and cohort shifts common in biomedicine (Ovadia et al. 2019). Uncertainty is most valuable when a system moves beyond familiar conditions, yet those conditions can also weaken its uncertainty estimates. A warning mechanism must therefore be evaluated under shift rather than assumed to generalize because it works on ordinary held-out data. Different shifts can affect ranking, calibration, and error concentration differently. No single in-distribution curve demonstrates protection against them all. This is methodological support for examining evidence independence, not a claim that the source studied immune-state and tumor evidence synthesis. A separate application-specific test is required at the interface described in this review.

Selective prediction formalizes abstention when evidence does not support action (Geifman and El-Yaniv 2017). Abstention creates a tradeoff between coverage and residual error. Comparing systems without matching coverage can make one look safer merely because it leaves more work unfinished. Deferred cases also consume human or computational resources. The evaluation should show the risk-coverage relation and account for the fate of excluded cases, rather than treating them as though they have ceased to matter. The connection to counting independent evidence in mechanistic reviews is a constraint on interpretation. It would be inappropriate to borrow the source's methodological authority as direct confirmation of the integrated domain claim.

Feature attribution can guide audit but does not establish biological causality (Lundberg and Lee 2017). Attribution decomposes a model output according to a defined explanatory convention. That decomposition can reveal dependence on suspicious features without establishing why the outcome occurs in the world. Reference choices and feature dependence can affect interpretation. A useful audit should compare explanations with known model behavior and should not transform a feature contribution into a treatment or mechanism claim. This is methodological support for examining evidence independence, not a claim that the source studied immune-state and tumor evidence synthesis. A separate application-specific test is required at the interface described in this review.

Health-AI governance requires accountability, transparency, inclusion, and protection of autonomy (World Health Organization 2021). Health-AI governance concerns people affected by a system as well as the organization operating it. Transparency without a route to challenge an output is incomplete. Autonomy and inclusion require attention to who can obtain follow-up, who is represented in evaluation, and who bears false alarms. These obligations remain even when an assistance tool is described as advisory rather than decisive. For the comparison involving shared cohorts, overlapping reviews, source lineage, 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 immune-state and tumor evidence synthesis,

this distinction is relevant to evidence independence. The proposed failure scenario is several summaries counted as independent experiments; 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. Within immune-state and tumor evidence synthesis, this distinction is relevant to evidence independence. The proposed failure scenario is several summaries counted as independent experiments; 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 evidence independence, not a claim that the source studied immune-state and tumor evidence synthesis. A separate application-specific test is required at the interface described in this review.

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 tissue contexts, clinical endpoints, and mechanistic claims. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

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 tissue contexts, clinical endpoints, and mechanistic claims. 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. An implementation should expose this limitation to reviewers of tissue contexts, clinical endpoints, and mechanistic claims. The result and its evidence qualification should remain recoverable even when the interface condenses them into a short report.

Research design and stopping conditions

Map datasets and overlapping samples before synthesis. Weight replication by independent data and design differences rather than by bibliography length. This would operationalize the concern with evidence independence 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 several summaries counted as independent experiments, with its validity judged independently of the system's response. The relevant comparison axes are shared cohorts, overlapping reviews, source lineage.

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 immune review, proposed mechanisms should be tested in their stated compartments and disease settings. Spatial association, state annotation, and intervention response would be judged separately. A useful result could expose a missing bridge in the integrated account. It would not need to confirm every part of the preferred narrative to improve scientific understanding.

What remains unverified

The cited sources still have design-specific limits even where their abstracts and methods are accessible. 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 independence changes the interpretation of shared cohorts, overlapping reviews, source lineage. 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 counting independent evidence in mechanistic reviews should end with a bounded claim: useful evidence can be preserved without pretending that every downstream inference has been tested. Map datasets and overlapping samples before synthesis. Weight replication by independent data and design differences rather than by bibliography length. The primary source remains bounded by its own design: A unifying set point is a hypothesis, not an already validated clinical measurement or an established therapeutic protocol. 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. “Resetting Immunometabolic Set Points in Autoimmune Disease.” *Journal of Translational Autoimmunity*, vol. 13, 2026, https://doi.org/10.1016/j.jtauto.2026.100391.

2. Liu, Sunjiang, et al. “Machine Learning-Based Prediction of Early Visual Field Nonrecovery after Endoscopic Surgery for Nonfunctioning Pituitary Adenomas with Clinically Accessible Biomarkers: A Dual-Center Study.” *International Journal of Clinical Practice*, vol. 2026, no. 1, 2026, https://doi.org/10.1155/ijcp/2504255.

3. Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

4. Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

5. Lundberg, Scott M., and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

6. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health*. World Health Organization, 2021.

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.