ASA Response to NIH Request for Information: Measuring and Rewarding Scientific Impact

August 19, 2026

Developed with scientific and technical input from Ryan A. Peterson, PhD, University of Iowa; Elizabeth W. Eisenhauer, PhD; Theresa Kim, PhD, MS; and other ASA members.

Claude Opus 5 (Anthropic) and GPT-5.6 Sol (OpenAI) were used to assist with drafting and editing these comments.

Contributor disclaimer: The views do not necessarily represent the views of contributors' employers or affiliated institutions. Theresa Kim provided scientific and technical input during off-duty hours and without use of government-furnished equipment; her participation does not represent the views of NIH, HHS, or the U.S. Government.

---

1. Rigor and Reproducibility

1.1 The missing instrument: indicators for analytic replicability

We are highly encouraged by NIH's renewed focus on transparency, reproducibility, and replication shown in the Request for Information on Measuring and Rewarding Scientific Impact, the Initiative to promote Strengthening Replication and Reproducibility of NIH-funded Research, the Gold Standard implementation plan, and the recent Highlighted Topic on Enhancing Scientific Rigor, Transparency and Replicability. These documents name reproducibility and replicability as distinct goals, correctly, and the

Initiative has instruments aimed at each: for the former, longstanding investments in data sharing, standardization, and public access to findings; for the latter, funded replication research through the Common Fund, a Replication Prize, and a stated intent to engage researchers in identifying areas ripe for replication.

These are steps in the right direction. However, we have yet to observe incentives targeting the gap between reproducibility and replicability: whether the reported inference is valid given how the analysis proceeded. Statisticians and biostatisticians have a useful and unusual vantage point on this gap: we are often among the first to recognize why a finding may not replicate, while not always having the standing or involvement needed to address the problem.

This matters because replication studies *detect* errors, but do not *prevent* them. They measure the failure rate after the fact, one finding at a time, at a high cost per finding. A study can be perfectly reproducible — fully containerized, one-click re-runnable — and still report inference that is unlikely to replicate, and the replication study will discover this several years and several hundred thousand dollars later.

A common route to this is inference on a model that was chosen using the same data used for inference: **unadjusted post-selection inference (UPSI)**. It is one instance of what Simmons, Nelson, and Simonsohn (2011) termed researcher degrees of freedom: undisclosed flexibility in analytic choices, sufficient on its own to produce significant findings without any intent to deceive. In a satirical blog post, Perry Hackman used simulations with pure-noise outcomes and a modest set of candidate subgroup interactions to show that an ostensibly rigorous method, forward stepwise selection with information criteria followed by standard statistical inference (dubbed UPSI), can produce at least one "significant" effect in 90+% of simulated trials (Peterson, 2025b). Nothing in that pipeline is clearly fraudulent; the analysis plan reads as rigorous, and the resulting code can be shared, containerized, and re-run perfectly.

Recommendation: The Initiative's replication instruments would return more per dollar if paired with a cheap upstream instrument that flags the findings most likely to need them. In analysis plans, progress reports, and publication reporting, a short structured statement answering the following could be incentivized:

- Was the reported model, subgroup, outcome scale, or covariate set chosen using the same data used to compute the reported inference?
- If so, how were the inferences adjusted for selection? (sample splitting; selective or conditional inference; bootstrap or model-averaged uncertainty; explicit relabeling as exploratory). If not, where was the full analysis specification verifiably pre-registered?
- How many candidate specifications were considered, including those examined and discarded?

Based on the above, results should be taken as on a spectrum between confirmatory (to provide stronger evidence for prespecified claims) and exploratory (to identify patterns and generate hypotheses).

Another recommendation is to commission periodic targeted statistical review of a random sample of supported publications, scored by trained statistical reviewers, reporting the proportion that (a) report selection-conditional inference without adjustment, (b) report subgroup findings without pre-specified subgroup structure, and (c) contain code that does not execute. Tracking that proportion over time measures whether the Initiative is working.

1.2 Candidate pools and model complexity are growing faster than the standards applied

Beneath many variable-selection procedures is an implicit assumption of covariate equipoise: that every candidate predictor is equally worthy of entering the model. Stepwise procedures, popular information criteria searches, and standard lasso implementations can all behave this way, and the assumption becomes problematic in settings where NIH is investing heavily.

Independent groups have converged on the same principle from different directions. Yu, Bien, and Tibshirani (2019) articulate a *reluctant interaction selection principle* — prefer a main effect over an interaction if all else is equal — and Tay and Tibshirani (2020) extend this method to a *reluctant non-linear selection principle*, preferring a linear term over a non-linear one on the same grounds. In his standard reference text, Harrell (2015) advises pre-specifying a small set of plausible interactions rather than testing all of them. Peterson and Cavanaugh (2022) advise that human-understandable "glass-box" models with meaningful features should be preferred to more complex ones precisely because they are easier to interrogate and replicate. Different penalties, different algorithms, same underlying claim: complexity should have to earn its way into a model, and treating every candidate term as equally deserving is what allows spurious complexity through. NIH need not adjudicate among these methods, but it should seek to require the reporting that makes the problem visible.

Transparent modeling methods matter because **replication requires something to replicate**: a claim specific enough that an independent team can test it on new data and report whether it held. Transparent statistical models often make such claims directly available. *Waist circumference is associated with one-year HDL, $\beta = -0.04$ (95% CI −0.07 to −0.01)* is a statement another group can go and check in a different dataset, and the check has a clear objective.

On the other hand, black-box predictive models, such as neural networks, XGBoost, random forests, and statistical models with high-order interaction terms, instead primarily yield prediction functions. A second team can apply the deposited model to new data and test whether predictive performance holds up. That is a meaningful form of external validation, but it is different from replicating an interpretable scientific claim. When performance drops, the failure may be difficult to localize: a transparent model can reveal which estimated relationship moved and in which direction, while an opaque model may reveal only that predictive performance deteriorated, and, if subgroup metrics are tracked, for whom.

This has a direct implication for the Initiative. Predictive modeling is an increasingly prominent category of NIH-supported research (e.g., Arshi et al. 2025, Le et al. 2024), and it is precisely the category on

which replication instruments have the least purchase. Transparency is therefore not only a matter of trust, fairness, or regulatory compliance, though it is those too. It also determines how specifically a scientific claim can be interrogated and replicated.

As datasets grow larger and contain newer, more high-dimensional modalities, new statistical methodology will be critical. For instance, with only twenty candidate predictors, to capture potential interactions, one would need to sift through 190 pairwise interactions, 1,140 three-way interactions, and over a million candidate models across all orders. Further, when modalities differ in their dimension and overall signal, they quickly run into practical challenges. For example, in a lung adenocarcinoma survival analysis with 7 clinical covariates and 22,283 gene expression measurements, assuming covariate equipoise and penalizing both groups equally encodes a structural disadvantage for clinically well-motivated variables, which get de-selected due to the sheer difference in dimensionality of the two modalities. With NIH's data-integration priorities, as studies become multimodal, consisting of genomics plus clinical plus imaging plus wearable plus environmental and more, the need for novel methods that can accommodate such data is critical; important methodological challenges remain at the scale and complexity of these data.

Recommendations: First, NIH should be wary of accepting post-hoc explanation of an opaque model (such as a locally interpretable model, a SHAP-based variable importance metric, etc.) as a reliable, replicable scientific finding; the explainer is a separate model that may or may not describe what the first one does, and should not be treated as equivalent to a directly interpretable scientific claim. Second, the accuracy cost of transparency is an empirical question that has been studied. A systematic review of 71 studies developing clinical prediction models found no performance benefit of machine learning over logistic regression (Christodoulou et al., 2019). Benchmarking work supports this outside the clinical setting as well: across 110 datasets from the Penn Machine Learning Benchmarks database, a transparent regression approach was within 5% of the best-performing method on 79% of classification and 62% of regression datasets, with random forests and gradient boosting showing no statistically distinguishable advantage (Peterson et al., 2024). Where transparent methods fell short, the reasons were diagnosable and specific and not reflective of a general law that complexity buys accuracy (specific examples can be found in Peterson et al., 2024, Section 3.3.2).

The implication is not that black-box methods should be discouraged, but that the burden of justification runs backward. **Investigators are rarely asked to show that their model's opacity bought anything**. For NIH-funded models intended to inform clinical decisions, a low-cost expectation is that a transparent baseline be fit and reported alongside the complex one, so the cost of interpretability is measured in that application rather than assumed. Where the gap is negligible, as is often the case, the transparent model is the better scientific product, because it can be interrogated and corrected.

1.3 Adequacy at the design stage: making evidentiary expectations auditable

These analysis-stage safeguards are necessary, but inference is only as strong as the data generated by the study design. The design determines which populations, outcomes, contrasts, and sources of variation are observed and therefore which scientific claims the resulting data can support. One of the earliest places this becomes explicit is the sample-size or design justification, where investigators state what they expect the proposed design to be capable of learning.

This makes the justification useful for more than determining a number. It provides a prospective, auditable statement of what adequate evidence was expected to look like. NIH can then ask whether the realized study provided that evidence and, if not, whether the departures and their implications were documented. A study need not unfold exactly as anticipated. Recruitment, measurement properties, effect sizes, intraclass correlations, or other design features must often be learned during the course of a study, at which point they become scientific information in their own right. Reporting them allows planning assumptions to be updated across studies rather than repeatedly regenerated from incomplete information.

For studies relying on existing or linked data, realized data quality may also differ from what was assumed at design, with implications for which claims the data can support. Rigor requires making those departures visible and reconsidering the claims they permit, rather than retrospectively treating the original justification as though nothing changed.

Gold Standard Science creates a policy demand for determining and justifying adequate sample sizes, but no common definition of “adequate” accompanies it. Raising conventional thresholds does not settle the problem. A study can have 90% power under poorly supported assumptions, for an effect of little scientific consequence, using an inefficient design, or with an imprecisely measured outcome. Adequacy is instead a fit-for-purpose judgment: whether the design, measurement, data quality, and analysis are sufficient for the claims the study is intended to support.

NIH guidance already reflects this principle. Under the current Parent R01 review framework, reviewers assess whether sample size is sufficient and well justified in relation to the research questions, outcomes, and design (NIH, 2024). NCCIH feasibility trials instead require quantitative feasibility or acceptability benchmarks and justification that the sample is sufficient to evaluate them; testing efficacy or “preliminary efficacy” is explicitly not the purpose (NCCIH, 2024a). By contrast, NCCIH efficacy, effectiveness, and pragmatic trials require power calculations linked to endpoints and hypotheses, at least 90% power for the primary outcome, adjustment for multiple primary outcomes, and a clinically meaningful effect size (NCCIH, 2024b). NIH therefore already treats adequacy as purpose-dependent; the opportunity is to make that principle explicit, consistent, and auditable.

The same principle extends beyond these examples. Exploratory secondary analyses, transportability or triangulation studies, resource construction, and methods development may require adequate precision, population overlap, analytic scope, ability to assess discordance, or support for future studies rather

than power against a single alternative. Requiring a power calculation regardless can produce the calculation without the justification. This is not a lower bar for exploratory work but a different one: identify the intended claim, the criterion of adequacy it requires, and the corresponding limits on inference.

Design information also interacts with the candidate analysis space. Limited information combined with broad analytic flexibility can generate apparently strong but unstable findings, while each source of risk is largely invisible to instruments aimed only at the other.

Declaring intended claims and criteria of adequacy at the design stage gives the disclosure in §1.1 something to be checked against. A confirmatory aim that later relies on substantial data-dependent selection, or a secondary outcome later presented as a primary claim, has changed evidentiary status during execution. That is not necessarily a scientific failure, provided that its path from "design" to "discussion" is visible and reflected in the claims made.

Recommendations:

First, require fit-for-purpose definitions of adequacy. Applications should identify the claims or objectives the study is intended to support, and the criterion for an adequate sample size or design should reflect that purpose. When an effect size is used to justify adequacy, its magnitude should be grounded where possible in empirical evidence and interpreted on a scale meaningful to the scientific question. Generic standardized effect-size conventions such as “small,” “medium,” and “large” should not substitute for separately considering the magnitude of effect that would matter and the variability anticipated in the study population. Not every aim of a complex study requires a separate, fully developed power analysis. However, the study should identify the minimum set of claims or objectives it is designed to support with strong evidentiary justification and demonstrate that the proposed design is adequate for those purposes.

Second, treat sample-size justification as an integral part of study design rather than a calculation performed after the design is largely fixed. The amount of information a study can provide depends not only on the number of observations but on interdependent choices about the population recruited, comparisons made, measurements collected, timing and intensity of follow-up, analytic strategy, and other features that can affect the credibility of conclusions. Those choices must also be made within constraints on participant burden, operational feasibility, time, and budget. Improving one dimension may worsen another. Sample-size justification should therefore be an iterative, collaborative process in which statistical, substantive, operational, and other relevant experts identify the major threats to the study’s success, examine how alternative design choices redistribute those threats, and determine how limited resources can be used to produce the strongest evidence for the study’s purpose. This provides a concrete, early example of collaboration that NIH could recognize and measure: integrating statistical, scientific, operational, and other relevant expertise into consequential study-design decisions.

Third, make sample-size and design justification transparent, auditable, and cumulative. Power analyses and other sample-size calculations should be reproducible from the statistical analysis plan, with the

method, inputs, assumptions, and software or tools documented sufficiently for an independent reviewer to recreate and vary the calculation. Studies should subsequently compare planning assumptions with what was realized, including recruitment and retention, variability, clustering, measurement performance, and other design quantities. Meaningful departures should be reported along with their implications for the evidence the study can support, rather than treated only as implementation failures. Recording what was expected, what occurred, and how those differences affected the study would allow empirical knowledge about design assumptions to accumulate across studies, so that future investigators can ground their assumptions in better evidence. NIH could track the proportion of funded studies that make their design justifications reproducible and subsequently report comparisons between key planning assumptions and realized study conditions.

1.4 Identifying findings ripe for replication

The Initiative states that NIH will engage researchers in selecting research areas ripe for replication and reproducibility studies. The indicators proposed above provide a practical way to do so.

Statistical fragility can triage more effectively than citation count alone. A tractable prioritization score could combine downstream influence (citations, guideline adoption, use as the basis for subsequent trials) with fragility markers extractable from the published record: limited effective information relative to analytic complexity; post-hoc subgroup findings; adaptive analytic choices; estimates where the magnitude just passes a significance threshold; single-site designs; and material departures from the design assumptions or adequacy criteria established at the outset. Findings that are both highly influential and structurally fragile are where replication dollars buy the most information.

The proposed indicators make these markers much easier to identify prospectively. Section 1.1 records whether inference was conditioned on data-driven selection; §1.3 records what the design was expected to support and whether realized information or study objectives materially changed. Issues like a lower-than-anticipated effective sample size, unexpectedly high or low intraclass correlation, or a claim that moved from exploratory to confirmatory status are not themselves a scientific failure. But when such changes materially affect the strength of a consequential finding, they provide a rational signal for replication priority.

1.5 Learning from discordance: reconciliation and transportability

Discordance in the existing evidence is itself informative. Studies that appear to address the same question may differ in estimand, population, measurement, design, or analytic assumptions. Triangulation uses this principle: agreement across approaches with different sources of bias can strengthen causal inference, while disagreement can reveal which assumptions or design features matter (Lawlor et al., 2016). Rather than defaulting to another *de novo* study, NIH could support structured reconciliation of discordant evidence, harmonizing these dimensions where possible and identifying the likely sources of remaining disagreement.

The same idea is more powerful when planned prospectively. Where transportability matters, a second cohort, site, health system, or data source can apply the same estimand, outcome definition, and prespecified analysis. This narrows the dimensions that vary and provides direct evidence about whether a result transports across populations or settings. Methods for generalizability and transportability already formalize this problem as extending causal inferences from a study population to a defined target population (Dahabreh et al., 2021).

One element of a definition of causal research is explicit definitions of the theoretical and empirical estimands and assumptions under which the design and analysis identify that quantity. Randomization alone does not remove this requirement: PCORI-funded work on post-randomization selection in cluster trials shows how treatment-dependent identification, recruitment, or noncompliance can change the population represented by the observed data and the causal effect that can be identified (Li et al., 2024).

Recent causal frameworks similarly emphasize the target estimand, target population, target trial, and target validity (Lu et al., 2024). Multiple data sources do not establish causality by themselves, but they can show whether a causal conclusion depends on the population, measurement, or contextual factors from which it was obtained.

Recommendations: Fund reconciliation of consequential discordant evidence as a scientific activity, with harmonized comparisons and identified sources of disagreement as deliverables rather than another narrative review. Where claims are intended to extend beyond a single source population, allow and incentivize budgets for harmonization and prespecified analysis in a second data source. The work required to make those sources comparable should itself be recognized as part of the scientific contribution.

1.6 A model that already works

NIH need not build a reproducibility review from scratch. Since 2016, the Journal of the American Statistical Association (JASA) has operated a reproducibility initiative establishing minimum criteria for code, data, and workflow, and created the Associate Editor of Reproducibility role to implement them; it has since expanded from Applications and Case Studies to all original research (Wrobel et al., 2024). Public materials include a reviewer and author guide, a pre-submission checklist, a standardized artifact declaration form, and a template repository.

Directly transferable elements:

- **A standardized artifact declaration**, submitted with progress reports or at closeout, stating what is reproducible, by which script, from which data, and where scope is limited and why.
- **A designated, credited reproducibility reviewer role.** The single most important design feature of the JASA program is that this is a named position with real editorial standing, not volunteer labor. NIH could support analogous roles in biostatistics cores as an allowable and expected budget line.

- **A recognition mechanism.** JASA gives an award for materials exceeding the minimum. NIH's Replication Prize already establishes the model; the same logic could extend to reproducibility artifacts, recognized in award reporting and citable in biosketches.
- **Master-script executability**: relative paths only, a wrapper script that reproduces results without modification, transparent run order, seeds set where exact reproduction is expected, versioned software documented and cited. Low-burden if adopted at project start rather than retrofitted.

---

2. Data, Software, and Model Sharing

Extend sharing from data to analytic artifacts. A dataset without the analysis is insufficient for validation. Deposited artifacts should include the executable pipeline, the computational environment (container or lockfile), seeds, and the *selection procedure itself,* assuming the model choice was data-dependent, not merely the final chosen model. Reporting only the winning specification discards what is needed to evaluate the inference.

Adopt synthetic facsimile data as the standard answer to restricted access. Where real data cannot be shared, JASA (for instance) requires a synthetic or facsimile dataset in the same format so code can at least be executed and inspected. This is a tested solution to NIH's hardest sharing problem and could be made an explicit, fundable expectation rather than an *ad hoc* accommodation.

Measure executed reproductions, not availability statements. The proportion of awards with a data availability statement is helpful but less meaningful as a quality signal. The meaningful metric is the proportion of deposited artifacts a qualified third party can execute end-to-end in a clean environment. NIH could seed this by funding a small number of institutional or consortium reproduction services and reporting aggregate success rates.

Fund data resources, not just data access. Valuable clinical and administrative data often require substantial legal, technical, and scientific work before they are usable beyond the institution that generated them. Curation, linkage, harmonization, missing-data assessment, and measurement-quality evaluation are part of evidence production, not background plumbing. NIH should fund and recognize this work, including contributions from the institutions and personnel that create usable data resources, and expect data quality to be assessed relative to specified use cases rather than treated as an intrinsic property of a dataset. PCORI's Patient-Centered Outcomes Data Repository provides one model: methodological projects can deposit documented, reusable study data expressly to support secondary analysis, reproduction, and further methods research (Li, 2026).

Integrating modalities is a methodological problem requiring renewed investment and awareness to protect the scientific record. Making heterogeneous data available together does not ensure that existing tools will handle them appropriately. Differences in measurement, missingness, dimensionality, and relevance to the scientific question across modalities require methodological development alongside

infrastructure investment. As argued in §1.2, funding data integration without funding the methods to analyze integrated data legitimately risks producing a great deal of accessible data and a great many non-replicable findings drawn from it.

Fund software maintenance. Methods software is infrastructure that decays without support. Statisticians and data scientists maintain packages on CRAN that are used well beyond the projects that produced them, with no funding attached to that maintenance — which is the ordinary arrangement rather than an unusual one. NIH should permit and expect maintenance budgets, count software citations as scholarly output, and expect critical software to be cited in references alongside papers so its use is traceable.

---

3. Training and Mentorship

Training impact should be measured as development, not placement alone. Career outcomes remain useful, but they depend on trainees' starting goals, opportunities, and the changing research labor market. NIH could instead pair placement outcomes with brief baseline and follow-up measures of trainees' skills, confidence, research responsibilities, and career plans. Useful indicators include the number of trainees supported; gains in competencies such as study design, data analysis, communication, collaboration, and reproducible practice; and whether trainees gained responsibilities or opportunities they did not have at entry.

Access to appropriate mentorship is itself measurable. Trainees on non-methodological projects should have documented access to statistical and other relevant quantitative mentorship with protected effort, and training plans should treat reproducible workflows, including version control, environment management, and executable analysis as basic research competencies rather than specialty skills.

Suggested indicators: baseline and follow-up assessments of trainee competencies and career plans; trainee placement interpreted relative to stated goals at entry; access to appropriate methodological mentorship; and inclusion of reproducible-research competencies in training plans.

---

4. Collaboration

Recognize methodological and reproducibility service as a scored contribution. Reproducibility review, cross-lab statistical mentoring, maintenance of shared analytic infrastructure, and other methodological service are largely invisible in conventional productivity metrics. Making these contributions visible in biosketches and review criteria is a low-cost way to recognize work that directly supports team science and research rigor. Independent methodological service, including statistical service on DSMBs and other

monitoring bodies, should likewise be recognized as a substantive contribution even when it does not result in co-authorship.

Collaborative statistical contribution is otherwise typically recorded as either co-authorship or an acknowledgment. Better indicators would capture both the role and structure of the collaboration, including the following:

- Named quantitative collaborators with **meaningful, specified effort**, ideally distinguishable from nominal statistical support at token (e.g., supervisory or consulting) levels.
- Contribution taxonomy (e.g., CRediT) at publication, naming methodology, software, formal analysis, and data curation roles.
- Collaborative products beyond publications, including jointly submitted grants, workshops or meetings developed, shared data or software resources, and trainees taught or mentored across disciplines.
- Whether the analysis plan was authored or co-authored by, and final analytic decisions authorized by, a named quantitative scientist with substantive responsibility for the analysis.

Collaboration should reflect more than the number of people a scientist has worked with, disciplines spanned, or projects contributed to. A quantitative scientist who develops a deep, sustained collaboration in one scientific area may contribute as much as one working across many fields. At the program or institutional level, NIH could instead characterize the reach and durability of collaborative networks, including repeated collaboration across projects and connections among otherwise separate scientific groups.

Analytic independence should be an explicit expectation. Individually reasonable analysis decisions, such as which covariates to adjust for, how to restrict a population, which subgroups to consider, how to handle outliers, etc., can cumulatively add up to an undocumented selection procedure, especially if there is pressure on such decisions in the direction of increased significance. Gelman and Loken (2014) call this the garden of forking paths: multiplicity exists even when only one analysis was ever run, because the path taken depended on the data. A quantitative collaborator supported only at token effort may lack the standing to resist result-driven analytic choices or to advocate independently for the ethical and judicious analysis.

Such independence should be supported by expecting named quantitative experts to participate meaningfully on research teams with substantive, not token, effort. Further, such independence is potentially measurable: whether the quantitative scientist holds authorship rather than acknowledgment, giving them standing to insist on a caveat or decline; whether their effort is protected on the award rather than discretionary; and, for high-stakes observational analyses, whether independent statistical review occurred, on the model of the independence already expected of trial statisticians.

5. Entrepreneurship and Translation

Translation measured through patents, licenses, and company formation systematically undercounts methodological translation. A method that moves from paper to open-source implementation to routine use in NIH-funded studies, clinical guidelines, regulatory submissions, or policy analysis is translated in the sense NIH cares about, and the process is measurable: downstream software citation, dependency counts in other NIH-supported analyses, adoption in guidelines and regulatory review.

NIH should include **methods-to-practice pathways** in translation indicators and support the unglamorous middle step between invention and adoption: testing, documentation, maintenance, training, and integration of methods into usable software and workflows. These activities currently fall through the cracks between methodological and translational funding.

---

6. Foundational Scientific Exploration

Valid inference after data-dependent selection is, at best, only partially solved. General solutions face fundamental limits; partial solutions exist but remain underused, in part because accounting honestly for selection often produces wider uncertainty, and wider uncertainty can be harder to publish.

Adjacent problems are similarly open and consequential: quantifying uncertainty for models produced by regularized selection; setting evidentiary thresholds that scale with the size and structure of a candidate space; establishing theoretical guarantees when covariate groups grow at different rates; and building transparent methods that remain competitive for outcome distributions where they currently underperform, such as zero-inflated and heavily skewed clinical endpoints.

This is precisely the foundational work the Initiative should fund: methodological, high-risk, without necessarily having an immediate clinical output, and directly relevant to whether findings across the rest of the portfolio replicate. As biomedical data and analytic choice expand, investment in replication alone addresses failures downstream; foundational statistical research can prevent some of them upstream.

---

7. Public Impact

Public impact should be evaluated along an impact pathway rather than inferred from citation counts alone. Evaluation frameworks distinguish intermediate outcomes such as reach, adoption, and sustained use from downstream changes in population health (Glasgow et al., 1999). NIH could similarly assess whether research produces trustworthy findings that reach intended populations, enter practice or policy, and are sustained over time.

Population outcomes are harder to attribute because the population is ultimately uncontrolled. Before asking whether research changed health outcomes, NIH often needs to know whether the resulting intervention, diagnostic, screening strategy, or clinical practice actually reached the population. NIH should incentivize studies linking changes in disease profiles and outcomes to changes in management and uptake, supported by data on where and among whom products and practices are used. This may require linking surveillance and health-system data with commercial, claims, registry, or other sources that capture diffusion beyond traditional research datasets.

Trustworthy evidence is a prerequisite for meaningful public impact. The replication rate of influential NIH-supported findings, tracked over time, would therefore provide an expensive but unusually direct indicator of whether the knowledge being propagated into research, practice, and policy remains supported when tested again.

Two additional indicators worth developing are the estimated resources committed downstream to findings that later failed to replicate, which NIH is uniquely positioned to assess; and the inspectability of clinical decision tools derived from NIH-funded research, including whether clinicians and patients can understand and contest the basis of the model's outputs.

Finally, not all public impact should be reduced to a count. Periodic expert evaluation of influential NIH-supported bodies of work could assess whether and how they changed scientific understanding, practice, policy, or population health over time.

---

8. Other Area: Generative AI, Agents, & Accountability

Generative AI and modern agentic analytic tools interact with every other category, and current trajectories run against the goals of this initiative.

AI-assisted analysis can quickly become an unrecorded selection procedure. An agent that iterates through dozens or hundreds of specifications and surfaces the one that "works best" is performing model selection at a scale exceeding what any analysis plan can reasonably specify in advance. The path taken by an LLM agent is rarely, if ever, reproducible from the prompt alone. If human-led data analysis represents Gelman's "garden of forking paths," an AI agent turns that garden into a random forest.

Mandating a complete audit trail of this search process is technically possible. However, with modern "chain of thought" LLMs, the massive volume of intermediate and erratic "thinking" steps renders any such record highly convoluted. An AI-assisted scientist must then decide among several alternatives. One option is to diligently store a large, complex record of these AI-dictated choices, and to publish this record alongside their research. Another option is to narratively summarize the AI-enabled search, which asks readers and funders alike to trust that whatever decisions were made, justified or not, were appropriate. Yet another option is to proceed as though the selected path was the "right" one all along.

Every incentive pushes the scientist toward this final path of least resistance, which also carries the worst consequences for replicability. Consequently, as AI-assisted analysis becomes normalized, rejected specifications and the rationale for moving among them will disappear — not because scientists intend to conceal them, but because we lack both the norms of analytic accountability and the practical, rigorous tools to support them. This is multiplicity at machine speed, and it is poised to undermine replicability if left unchecked.

Recommendation: prevent multiplicity at the design stage through methodological collaboration. Methodological input on study design should be a fundable, expected component of research, not an unbudgeted courtesy. It has always been true that a specification chosen *a priori* on a firm statistical foundation generates no search to document. Until recently, the effort required to explore many specifications served as an informal check on doing so. AI agents remove that friction, making design-stage input more consequential, yet simultaneously easier to skip. The recommendations that follow address the multiplicity that remains, as it always will, since no analysis plan can anticipate everything a real dataset will demand.

Recommended indicator: an analytic provenance record. Where generative AI materially contributes to analytic code or decisions, require a short log — model and version, date, nature of the contribution, and, where AI was used to generate or compare analytic alternatives, the alternatives considered and the selection process. Systems used for analytic work should make this information exportable rather than requiring investigators to reconstruct it after the fact. NIH could support development of provenance-capture tooling as research infrastructure.

Recommendation: retain locatable human accountability. A named individual should be able to explain and defend every analytic choice. NIH could require an attestation, parallel to existing authorship and conflict attestations, that a responsible analyst reviewed and understood AI-generated analytic code and decisions. AI may propose analytic choices, but a named human must adopt them and be able to state why each was selected over considered alternatives. This shifts analytic authority from AI dictation, which cannot be held to account for its mistakes, back to human judgment.

Recommendation: promote team science through shared analytic artifacts. Collaboration requires a shared, inspectable artifact to collaborate on. If each member works in a private AI-assisted workflow producing unshared code and unlogged decisions, team science becomes several parallel solo efforts with a joint author list (or worse, reverts to solo work with an AI). Shared repositories, executable master scripts, versioned environments, and declared selection procedures are the practical antidote, and NIH should fund and expect them at the team level.

9. Other Area: Unintended Consequences and Feasibility

- **Checklist theater.** Any requirement can become a compliance artifact if checked for existence rather than content; data-availability statements are a cautionary example of the limitations of

existence-based metrics. Mitigation: sample-based audits by qualified reviewers, and metrics based on execution rather than deposit.

- **Chilling exploration.** Pre-specification requirements can be misread as prohibiting exploration. The requirement is *labeling*, not prohibition. Selection disclosure should let exploratory work proceed freely and be reported honestly as exploratory rather than presented as confirmatory.
- **Over-rewarding confirmation.** If replication success becomes a scored metric for individual investigators, the rational response is to study safe questions. This ultimately risks stifled innovation and hinders research exploration. Replication incentives should attach to the enterprise and to replication teams, not to the original investigator's record.
- **Regressive burden.** Container registries, reproducibility reviewers, and synthetic data generation are easier to absorb at well-resourced institutions than at small ones. Mitigation: fund shared infrastructure, allow these costs explicitly in budgets, and scale expectations to award size and complexity rather than applying a uniform standard.
- **Career-stage asymmetry.** Compliance and service expectations impose different burdens across career stages, with early-career investigators often least able to absorb them. Reproducibility service and exemplary artifacts should count concretely in review.
- **Metric capture generally.** The indicators emphasized here, including executed reproductions, audited selection disclosure, candidate-space reporting, and downstream enablement, are relatively resistant to gaming because satisfying them requires producing much of the underlying scientific behavior they are intended to measure.

Feasibility across contexts: selection disclosure, candidate-space reporting, and artifact declaration are relatively low burden when incorporated prospectively into the workflow, because they record decisions as they occur rather than reconstructing them later. Executable-artifact standards carry moderate costs and scale with award size. Independent reproduction services are more resource-intensive and should be centrally or consortium-funded rather than mandated institution by institution.

Closing

Statisticians and biostatisticians have an unusual vantage point on this RFI: we are often among the first to recognize why a finding may not replicate, while not always having the standing or involvement needed to address the problem. The interventions most likely to succeed make analytic choices visible and attributable, because visibility is what allows a collaborator, reviewer, or reader to catch the problem, and because the alternative in an era of AI-accelerated analysis is a scientific record that is easy to re-run and impossible to interrogate.

The American Statistical Association appreciates the opportunity to comment on the Request for Information on Measuring and Rewarding Scientific Impact. Questions or comments may be directed to ASA Director of Science Policy Steve Pierson: spierson@amstat.org.

References cited

- Arshi, B., Wynants, L., Rijnhart, E., Reeve, K., Cowley, L. E., & Smits, L. J. (2025). Number of Publications on New Clinical Prediction Models: A Bibliometric Review. *JMIR Medical Informatics*, *13*, e62710. https://doi.org/10.2196/62710
- Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004
- Dahabreh, I. J., Haneuse, S. J. A., Robins, J. M., Robertson, S. E., Buchanan, A. L., Stuart, E. A., & Hernán, M. A. (2021). Study Designs for Extending Causal Inferences From a Randomized Trial to a Target Population. *American Journal of Epidemiology*, *190*(8), 1632–1642. https://doi.org/10.1093/aje/kwaa270
- Gelman, A., & Loken, E.R. (2014). The Statistical Crisis in Science. *American Scientist*, 102, 460. https://doi.org/10.1511/2014.111.460
- Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the Public Health Impact of Health Promotion Interventions: The RE-AIM Framework. *American Journal of Public Health*, *89*(9), 1322–1327. https://doi.org/10.2105/ajph.89.9.1322
- Harrell, F.E. (2015). *Regression Modeling Strategies*, 2nd ed. Springer.
- JASA Reproducibility Guide and Pre-Submission Author Checklist: https://jasa-acs.github.io/repro-guide/
- Lawlor, D. A., Tilling, K., & Davey Smith, G. (2016). Triangulation in Aetiological Epidemiology. *International Journal of Epidemiology*, *45*(6), 1866–1886. https://doi.org/10.1093/ije/dyw314
- Le, J. P., Morrison, J., Malhotra, A., Nemati, S., Wardi, G., & Ford, J. S. (2026). National Institutes of Health-Funded Artificial Intelligence and Machine Learning Research, 2019-2023: Cross-Sectional Study. *Journal of Medical Internet Research*, *28*, e84861. https://doi.org/10.2196/84861
- Li, F., Li, F., Liu, B., Papadogeorgou, G., Bobb, J., Thomas, L., & Wruck, L. (2024). *New Causal Inference Methods for Cluster-Randomized Trials with Postrandomization Selection Bias*. Patient-Centered Outcomes Research Institute (PCORI). https://doi.org/10.25302/10.2024.ME.2019C116146
- Lu, H., Li, F., Lesko, C. R., Fink, D. S., Rudolph, K. E., Harhay, M. O., Rentsch, C. T., Fiellin, D. A., & Gonsalves, G. S. (2024). Four Targets: An Enhanced Framework for Guiding Causal Inference from Observational Data. *International Journal of Epidemiology*, *54*(1), dyaf003. https://doi.org/10.1093/ije/dyaf003
- National Center for Complementary and Integrative Health. (2024a). *PAR-25-274: Feasibility Clinical Trials of Mind and Body Interventions for NCCIH High Priority Research Topics (R34 Clinical Trial Required).* National Institutes of Health. Posted November 21, 2024. Accessed August 17, 2026.
- National Center for Complementary and Integrative Health. (2024b). *PAR-25-268: Investigator Initiated Clinical Trials of Complementary and Integrative Interventions Delivered Remotely or via*

mHealth (R01 Clinical Trial Required). National Institutes of Health. Posted November 21, 2024. Accessed August 17, 2026.

- National Institutes of Health. (2024). *PA-25-305: NIH Research Project Grant (Parent R01 Clinical Trial Required).* Posted December 18, 2024. Accessed August 17, 2026.
- Peterson, R.A., & Cavanaugh, J.E. (2022). Ranked Sparsity: A Cogent Regularization Framework for Selecting and Estimating Feature Interactions and Polynomials. *AStA Advances in Statistical Analysis*, 106, 427–454. https://doi.org/10.1007/s10182-021-00431-7
- Peterson, R.A., McGrath, M., & Cavanaugh, J.E. (2024). Can a Transparent Machine Learning Algorithm Predict Better than Its Black Box Counterparts? A Benchmarking Study Using 110 Data Sets. *Entropy*, 26, 746. https://doi.org/10.3390/e26090746
- Peterson, R.A. (2025a). What Do We Mean by *Glass-box*, Exactly? *Data Diction (blog)*. https://doi.org/10.59350/9dhes-thd51
- Peterson, R.A. (2025b) [written under Hackman, P. pseudonym]. How Can I Guarantee a Significant Result? *Data Diction (blog)*. https://doi.org/10.59350/6mzf8-xzd69
- Peterson, R.A., Bird, S.M., Harris, L.M., Breheny, P.J., & Cavanaugh, J.E. (2026). A Ranked Sparsity Extension to the Bayesian Information Criterion: A Tool for Selecting Variables from Multiple Data Modalities. Preprint (currently in resubmission to *Entropy*). https://doi.org/10.20944/preprints202607.0193.v1
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632
- Tay, J.K., & Tibshirani, R. (2020). Reluctant Generalised Additive Modelling. *International Statistical Review*, 88(Suppl 1), S205–S224. https://doi.org/10.1111/insr.12429
- Wrobel, J., Hector, E.C., Crawford, L., D'Agostino McGowan, L., da Silva, N., Goldsmith, J., Hicks, S., Kane, M., Lee, Y., Mayrink, V., Paciorek, C.J., Usher, T., & Wolfson, J. (2024). Partnering with Authors to Enhance Reproducibility at JASA. *Journal of the American Statistical Association*, 119(546), 795–797. https://doi.org/10.1080/01621459.2024.2340557
- Yu, G., Bien, J., & Tibshirani, R. (2019). Reluctant Interaction Modeling. Preprint. https://doi.org/10.48550/arXiv.1907.08414