PRE-VALIDATION WORKING PAPER

Toward Objective and Auditable Data Analysis

The GHEAS Auditable Evidence Framework for Design Credibility, Measurement Reliability, and Robustness

Everion Cao

GHEAS, Waterloo, Ontario, Canada

ORCID 0009-0005-8238-8583 | everioncao@gmail.com

Methodological blueprint v1.0 | 21 July 2026

Abstract

Background

Statistical analyses can be technically correct yet produce fragile or misleading conclusions because the research question is underspecified, data provenance is incomplete, measurements are unreliable, several reasonable analytic choices exist, or causal language exceeds the study design. Selective reporting is only one part of the problem. A credible workflow must also expose how data construction, model specification, temporal alignment, measurement error, missingness, and falsification results affect the conclusion.

Objective

To specify a domain-adaptable method and software contract for producing statistical conclusions that are more transparent, impartial, reproducible, and appropriately calibrated to the available evidence.

Framework

The GHEAS Auditable Evidence Framework (GAEF) treats an analysis as a versioned evidence process rather than a preferred model. It contains nine linked components: (1) a claim and estimand contract; (2) a data provenance and measurement contract; (3) a design and variable-role map; (4) separate associational and identification tracks; (5) a prespecified universe of reasonable analyses; (6) falsification and failure gates; (7) quantitative bias and uncertainty analysis; (8) evidence reconciliation with rule-bound claim licensing; and (9) an executable audit trail with bounded AI assistance. The framework permits "insufficient evidence" and "design failure" as valid outputs and prohibits a statistically significant result from automatically licensing a causal claim.

Validation plan

GAEF will be evaluated sequentially using software unit tests, simulations with known truth, adversarial data perturbations, public benchmark data, a real-world PM2.5 and health case study, and at least one cross-domain application. Primary evaluation criteria include estimand recovery bias, interval coverage, false claim rate, fault-detection performance, robustness-envelope stability, computational reproducibility, and audit completeness.

Conclusion

This paper is a specification, not evidence that GAEF is superior to existing practice. Its central proposition is testable: a workflow that records provenance, exposes reasonable analytic alternatives, requires falsification, and gates claim strength should reduce unsupported conclusions without hiding genuine uncertainty. Only the validation programme can determine whether that proposition is correct.

Keywords: data analysis; reproducibility; robustness; causal inference; measurement error; multiverse analysis; artificial intelligence; audit trail

1. Introduction

The visible output of a statistical study is often a coefficient, interval, prediction, or graph. The conclusion, however, depends on a longer chain of decisions: what question was asked, which population and contrast were targeted, how raw observations were produced, which records were retained, which variables were treated as confounders or consequences, how time was aligned, which model was selected, and what evidence was allowed to contradict the preferred interpretation.

Conventional reporting can conceal this chain even when no misconduct occurs. Analysts may face many defensible choices in data construction and modelling. A single selected analysis cannot show whether the conclusion is stable across those choices. Multiverse and specification-curve methods were developed to make such dependence visible [4,5]. Likewise, a small p-value does not measure effect size, practical importance, or the probability that a scientific claim is true [1]. These limitations are not repaired merely by using a more complex model or an artificial-intelligence system.

Other methodological traditions address different parts of the chain. Estimand frameworks require the target of estimation to be defined before the estimator is chosen [2]. Target-trial thinking aligns eligibility, treatment or exposure assignment, time zero, follow-up, outcomes, and analysis [3]. Negative controls can reveal noncausal associations that remain after adjustment [6]. Quantitative bias analysis can evaluate how measurement, confounding, or selection error might change a result rather than relegating those errors to a limitations paragraph [7]. FAIR data principles and reproducible-computing guidance emphasize persistent identifiers, machine-readable metadata, complete records of transformations, and retention of intermediate results [8,9].

These practices are valuable but are commonly implemented as separate recommendations. GAEF is intended to join them into one executable contract. Its purpose is not to produce a universal statistical model. It is to control the relationship among question, data, design, analysis, failure tests, uncertainty, and language.

The framework emerged from development work on ambient particulate pollution and population health. In that work, apparently interpretable associations changed under country-specific trends and temporal placebo tests, while three candidate policy designs were stopped before health-outcome estimation because their pollution first stage, pre-trends, or trend robustness did not satisfy the prespecified gate. Those experiences are motivating examples, not a validation of GAEF. A method must be tested where truth or injected error is known before success in a complex observational case can be interpreted.

2. Aim and scope

GAEF is designed for quantitative analyses that use observational, experimental, administrative, environmental, laboratory, or linked data. It can support descriptive, associational, predictive, or causal questions, but the requirements differ by claim class. The first implementation focuses on population-health and environmental data; the conceptual contract is intended to be portable.

The framework has three aims:

1. reduce unrecorded analytic discretion;
2. detect conditions under which a numerical result should not support the intended claim; and
3. make the complete path from source data to wording independently auditable.

GAEF does not guarantee truth, remove the need for subject-matter knowledge, or make all reasonable analysts agree. It cannot recover information that was never measured, identify an effect without defensible variation, or prove that unmeasured bias is absent. It is designed to make those boundaries explicit and consequential.

3. Operational definitions

3.1 Accuracy

Accuracy is closeness to a defined target, not agreement with a published answer. When the data-generating process is known, accuracy can be evaluated by bias, root mean squared error, interval coverage, calibration, and error rates. In real observational data where truth is unknown, the framework reports identification assumptions, measurement validity, robustness, and uncertainty; it does not relabel consistency as truth.

3.2 Objectivity

Objectivity is treated as a property of the process. A process is more objective when important choices are explicit, rules are applied before their consequences are known, reasonable alternatives are examined symmetrically, and another analyst can reconstruct how the conclusion was licensed. Preregistration can help distinguish prediction from post hoc explanation, but registration alone does not guarantee a sound design or correct analysis [10].

3.3 Impartiality

Impartiality means that evidence supporting and opposing a claim is subjected to the same inclusion, diagnostic, and reporting rules. A null result is not automatically privileged over a non-null result, and statistical significance does not exempt a model from falsification. Failed designs and contradictory models remain in the record.

3.4 Auditability

Auditability means that an independent reviewer can trace every key number and claim to a versioned input, transformation, model, diagnostic, decision rule, and software environment. Reproducibility is necessary but not sufficient: a perfectly reproducible analysis can still answer the wrong question.

3.5 Calibrated conclusion

A conclusion is calibrated when its wording does not exceed the strongest claim class supported by the design and diagnostics. "The data describe," "the variables are associated," "the model predicts," and "the design supports a bounded causal interpretation" are distinct outputs.

4. The GHEAS Auditable Evidence Framework

4.1 Component 1: claim and estimand contract

Each project begins with a machine-readable contract specifying:

- the scientific question and claim class;
- target population, unit, exposure or intervention, comparator, outcome, and time horizon;

- the population-level summary and weighting scheme;
- the theoretical target, empirical estimand, estimator, and conditions linking them;
- the minimum evidence required to license the intended language; and
- stopping conditions that make the target unidentifiable with current data.

The estimand is fixed before a preferred estimator is selected. Different weighting schemes, exposure increments, time horizons, or outcome definitions are distinct estimands, not interchangeable sensitivity analyses. The contract can be amended, but amendments are dated, justified, and labelled according to whether relevant results were visible.

4.2 Component 2: data provenance and measurement contract

Every input must identify its provider, version or access date, licence, geographic and temporal coverage, unit, measurement type, linkage method, and known processing history. Cryptographic hashes preserve file identity. Raw files are immutable; derived files are produced by code.

The measurement contract distinguishes direct observations, instrument readings, administrative records, modelled estimates, imputations, and hybrid measures. It records calibration, blanks, replicates, batch, laboratory, inter-rater, linkage, and validation information when relevant. The absence of public raw data is recorded as limited verifiability, not proof of error.

Unit consistency is a formal gate. A value measured as PM10, for example, cannot be relabelled as PM2.5. A conversion requires an explicit measurement model, an evidence source for the conversion parameters, and propagation of conversion uncertainty.

4.3 Component 3: design and variable-role map

Variables are assigned scientific roles before model fitting: exposure, outcome, confounder, mediator, collider or descendant, precision variable, negative control, instrument, effect modifier, or descriptive variable. Timing is part of the role. A variable measured after exposure or affected by the outcome cannot be treated as an ordinary baseline confounder without a stated causal rationale.

The design map records selection into the analytic sample, measurement processes, interference, missingness, temporal ordering, and the source of identifying variation. A directed acyclic graph may support this task but does not replace the written assumptions.

4.4 Component 4: separate evidence tracks

GAEF separates analyses by the claims they can support.

Track A: strengthened observational analysis. Track A estimates descriptive or conditional associational quantities. It requires the unadjusted estimate, prespecified adjustment sets, appropriate spatial or unit and time structure, alternative exposure windows, missingness analyses, reasonable trend controls, and uncertainty estimators. Track A can show that an association is stable or fragile; it cannot become causal merely because multiple models agree.

Track B: explicit identification analysis. Track B requires a randomized assignment mechanism or a documented quasi-experimental source of variation, such as a threshold, policy timing, boundary, instrument, or natural event. The analysis must define the counterfactual contrast and test design-specific implications, including exposure first stage, pre-trends or continuity, anticipation, spillovers, treatment reversal, and negative controls. If the identification gate fails, the health-effect estimate remains deliberately empty.

Tracks are not mechanically averaged. Agreement improves coherence; disagreement triggers an investigation of population, time, measurement, estimand, and assumptions.

4.5 Component 5: reasonable-analysis universe

Before the main result is interpreted, the project defines the set of analyses that are both scientifically defensible and materially relevant. Dimensions may include sample eligibility, variable construction, exposure windows, missing-data handling, adjustment sets, functional form, weighting, dependence structure, and uncertainty estimator.

The aim is not to run every mathematically possible model. Each branch requires a rationale, and logically invalid branches are excluded before results are examined. Outputs include a complete specification table, distribution of estimates, sign and interval stability, influential choices, and the subset that satisfies all diagnostic gates. This extends multiverse and specification-curve logic by connecting each model to claim licensing rather than treating model frequency as a vote on truth [4,5].

4.6 Component 6: falsification and failure gates

GAEF requires analyses that could make the preferred interpretation less credible. Depending on design, these may include:

- future-exposure and impossible-time placebo tests;
- negative-control outcomes or exposures [6];
- pre-trend, anticipation, continuity, balance, or manipulation checks;
- synthetic null data and outcome permutation;
- unit, time, coding, and sample-flow consistency checks;
- out-of-sample or temporal validation for prediction;
- leave-one-unit, leave-one-cluster, or influence analyses; and
- benchmark reproduction under an independently implemented estimator.

Failure is not automatically proof that the substantive effect is absent. It is evidence that the tested design or model does not license the intended claim. The action attached to each gate—continue, downgrade, quarantine, or stop—is specified in advance.

4.7 Component 7: bias and uncertainty analysis

Sampling uncertainty is only one component of uncertainty. GAEF distinguishes:

- random error;
- measurement error and misclassification;
- missingness and selection;
- uncontrolled or poorly measured confounding;
- dependence and clustering;
- model and functional-form uncertainty;
- data-version and linkage uncertainty; and

- transportability uncertainty.

Where information permits, deterministic or probabilistic quantitative bias analysis propagates plausible error parameters through the estimate [7]. A measurement-error correction is not accepted solely because it increases an effect. Its assumptions, validation source, and sensitivity to mis-specification must be shown.

4.8 Component 8: evidence reconciliation and claim licensing

The framework produces an evidence ledger rather than one privileged p-value. For each claim it records supporting evidence, contradictory evidence, unresolved assumptions, diagnostic failures, and the result's dependence on analytic choices.

The reference claim ladder is:

1. E0 — ungradable: data, measurement, or design is insufficient;
2. E1 — descriptive: a pattern is documented in a defined dataset;
3. E2 — associational: an association persists across prespecified reasonable analyses;
4. E3 — robust associational or predictive: diagnostics, validation, and bias analyses support a stable noncausal or predictive claim; and
5. E4 — bounded causal: an explicit identification design passes its prespecified gates and supports a causal interpretation under stated assumptions.

The ladder is not a quality score. A strong E2 association and an inconclusive E4 attempt answer different questions. AI cannot promote a claim level.

4.9 Component 9: executable audit trail and bounded AI

Every run produces a manifest containing protocol hash, input hashes, software versions, random seeds, sample flow, model inventory, diagnostics, exclusions, amendments, and generated outputs. Intermediate results are retained, following the principle that reproducible computation requires complete records of how each result was produced [9]. Data and metadata should be findable, accessible under appropriate permissions, interoperable, and reusable where possible [8].

AI may assist with code generation, data-quality checks, literature triage, model execution, comparison, and report drafting. It may not silently alter the question, delete contradictory models, invent missing provenance, adjudicate high-stakes exclusions without review, or upgrade causal language. AI outputs must be attributable to a model and session or workflow version and remain subject to deterministic tests and human responsibility. This bounded role is consistent with broader expectations that AI systems be valid and reliable, accountable and transparent, and evaluated throughout their lifecycle [13].

5. Reference software architecture

The framework maps to seven software layers:

1. Protocol layer: JSON schema and human-readable study contract.
2. Data registry: source, licence, measurement, coverage, and hash records.
3. Validation layer: schema, unit, range, key, timing, missingness, and duplication checks.

4. Analysis engine: claim-specific Track A and Track B estimators plus the reasonable-analysis universe.
5. Falsification engine: negative controls, placebo tests, design gates, influence analyses, and adversarial checks.
6. Evidence reconciler: structured comparison and rule-bound claim level.
7. Report and audit generator: complete tables, figures, limitations, manifests, and machine-readable results.

The system must fail closed: if a required input, assumption, diagnostic, or provenance field is missing, it should not silently continue to a stronger claim. A warning must change the analysis state, not merely appear in a log.

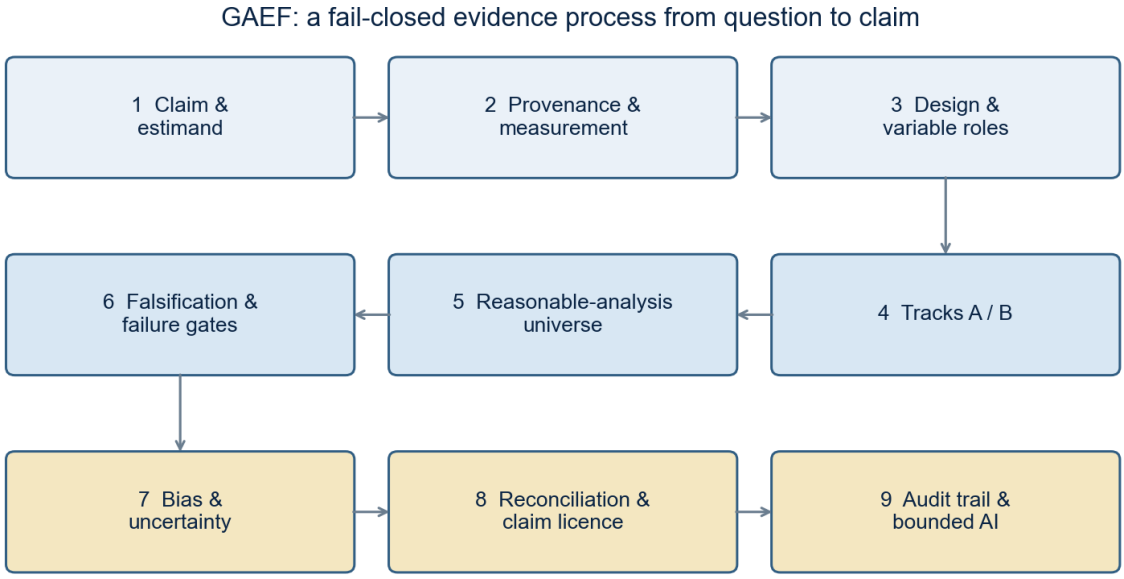

Figure 1. GAEF links specification, analysis, falsification, and claim licensing in a fail-closed workflow.

6. Validation programme

6.1 Stage V1: software correctness

Unit tests will verify transformations, estimators, standard errors, confidence intervals, hashes, and decision rules. Deterministic fixtures will test known null, positive, negative, and heterogeneous effects. Independent implementations will be compared within numerical tolerance. Malformed inputs must trigger the intended failure state.

6.2 Stage V2: simulation with known truth

Factorial simulations will vary sample size, effect magnitude, nonlinearity, heterogeneity, clustering, serial correlation, exposure persistence, confounding strength, interference, and policy adoption. Scenarios will include zero effects and sign reversals. Primary metrics are bias, root mean squared error, 95% interval coverage, type-S error, type-M error, and false claim-level promotion.

The simulation code and evaluation criteria will be fixed before aggregate performance is inspected. A method will not pass merely because it performs well in the data-generating processes for which it was designed.

6.3 Stage V3: adversarial perturbation

Clean datasets will be modified by injecting known faults: unit conversion errors, duplicate rows, shifted timestamps, laboratory batch effects, exposure measurement error, outcome leakage, missing-not-at-random selection, coding changes, extreme leverage points, and selective removal of models. Detection sensitivity, false alarm rate, correct severity classification, and resulting claim downgrade will be measured.

6.4 Stage V4: public benchmarks

Public datasets with documented analyses or known design properties will be reanalysed. Agreement with the published result will be treated as a reproducibility outcome, not as proof that the published result is correct. Where possible, analysts assessing the final output will be blinded to the benchmark result until the protocol and code are locked.

6.5 Stage V5: PM2.5 and health case study

The existing GHEAS assets provide three complementary tests:

- a country-year PM2.5 and life-expectancy panel for model dependence and temporal falsification;
- audited policy or natural-experiment candidates for Track B failure gates; and
- a long-term PM2.5 and mortality evidence corpus for overlapping studies, exposure measurement, risk of bias, and evidence synthesis.

Traditional analysis, Track A, Track B, and the final evidence ledger will be compared after the framework implementation is frozen. The case study will not be used to tune rules and then presented as independent validation.

6.6 Stage V6: cross-domain transfer

At least one application outside air pollution will be selected before final claims of generality. Candidate domains include water contamination, climate and health, nutrition, laboratory measurement, or another longitudinal public dataset. Domain-specific modules are permitted; the core claim and audit contract must remain unchanged.

7. Evaluation and acceptance criteria

Before validation results are available, this paper prespecifies the following families of criteria:

- Numerical validity: estimator bias, error, interval coverage, and calibration under known truth.
- Decision validity: rate of inappropriate causal or directional claims, including under null and confounded scenarios.
- Fault detection: sensitivity, specificity, and severity classification for injected data and workflow errors.
- Robustness transparency: proportion of defensible specifications shown, choices driving material changes, and completeness of contradictory results.

- Reproducibility: identical manifests and results under the same inputs and environment, and bounded tolerance across independent implementations.
- Audit completeness: proportion of key claims traceable to source, transformation, model, diagnostic, and rule.
- Usability: time, expertise, and human intervention required; frequency of warnings that are technically correct but not actionable.

Exact pass thresholds will be fixed in the validation protocol before final testing. This prevents the blueprint from declaring success using thresholds chosen after performance is known.

8. Relationship to existing methods

GAEF does not replace domain-specific statistical methods or reporting guidelines. It incorporates and connects established ideas:

- explicit estimands and alignment of design, analysis, and interpretation [2,3];
- multiverse and specification-curve analysis [4,5];
- negative controls and quantitative bias analysis [6,7];
- FAIR data stewardship and reproducible computation [8,9];
- separation of preregistered and exploratory work [10];
- institutional recommendations for reproducibility and replicability [11,12]; and
- lifecycle accountability for AI-supported systems [13].

The proposed contribution is an executable integration with claim-level gates, not invention of those component methods. Domain reporting standards remain necessary. For example, a prediction-model study should still use applicable transparent reporting guidance such as TRIPOD+AI [14].

9. Governance and human responsibility

The phrase "AI does the analysis" can obscure responsibility. Under GAEF, humans remain responsible for the scientific question, ethical boundaries, variable roles, causal assumptions, access permissions, interpretation, and release. Data constrain what can be observed. AI and software execute declared rules, identify inconsistencies, compare alternatives, and document the process.

All consequential overrides require a reason and identity. Manual changes are not prohibited; unrecorded changes are. An appeal mechanism should permit a human reviewer to challenge a gate, but the original decision and the override must both remain visible.

Privacy and fairness require domain-specific assessment. An auditable model can still be harmful, unrepresentative, or used outside its intended population. GAEF therefore distinguishes statistical impartiality from social fairness and does not claim that one guarantees the other.

10. Limitations and possible failure modes

The framework may create excessive complexity, false reassurance, or a large volume of diagnostics that users cannot interpret. Prespecification can freeze poor assumptions. A reasonable-analysis universe can itself be biased by what its designers regard as reasonable. Failure gates may be underpowered, and passing a negative control does not prove absence of confounding. Quantitative bias analysis depends on uncertain bias parameters. Reproducible code can reproduce systematic error.

Automation introduces additional risks: incorrect code may be executed at scale; AI-generated explanations may sound more certain than the evidence; software defaults may become hidden methodological authority; and users may treat an evidence grade as a substitute for judgment. These are reasons to test the framework adversarially and to report its own errors.

The initial development team and case study are limited. Independent users, domains, and implementations will be required before broad claims of validity or generality are justified.

11. Discussion

The central design choice in GAEF is to make a conclusion the output of an evidence process rather than the verbal interpretation of one model. The process can end with a number, a bounded claim, a downgraded claim, or a stop. Treating a stop as a legitimate result changes the incentive structure: an analysis does not have to produce a health-effect estimate to be useful.

This approach also changes the role of a paper. The present manuscript is a specification against which software and validation can be tested. It will be revised after failures are observed. A later paper may report successful and unsuccessful validation results, but publication is not the acceptance test. The acceptance test is whether the method performs as prespecified under known truth, detects meaningful faults, preserves contradictory evidence, and calibrates claims in real applications.

The PM2.5 example is valuable precisely because it is difficult. Exposure is modelled, populations and time scales differ, confounding is substantial, policies may have weak first stages, and prior results are heterogeneous. It is therefore a strong stress test for usability and auditability, but a poor sole test of statistical accuracy because the true population effect is unknown.

12. Conclusion

GAEF proposes a method-first route to more objective and auditable data analysis. It does not ask one model, one analyst, or one AI system to be unbiased. It constrains the process through explicit targets, provenance, measurement assessment, reasonable alternatives, falsification, quantitative uncertainty, claim gates, and a complete audit trail.

The framework remains a hypothesis. The next task is not to promote it but to try to break it using known-truth simulations, adversarial errors, public benchmarks, and difficult real data. A method that cannot survive those tests should be revised or rejected, regardless of how persuasive its paper appears.

Declarations

Funding: This work received no external funding.

Conflicts of interest: The author declares no conflicts of interest.

Data and code availability: The working framework, protocols, code, tests, and case-study assets are maintained in the versioned GHEAS project. Public release will follow validation and privacy/licence review.

AI assistance: OpenAI Codex, referred to in the project as Jace, assisted with literature organization, drafting, code development, and document production under human direction. It is not an author and does not assume responsibility for the work. All scientific claims and release decisions remain the responsibility of the named author.

References

1. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. The American Statistician. 2016;70(2):129-133. https://doi.org/10.1080/00031305.2016.1154108
2. International Council for Harmonisation. ICH E9(R1): Addendum on estimands and sensitivity analysis in clinical trials. 2019. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf
3. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. Journal of Clinical Epidemiology. 2016;79:70-75. https://doi.org/10.1016/j.jclinepi.2016.04.014
4. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science. 2016;11(5):702-712. https://doi.org/10.1177/1745691616658637
5. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. Nature Human Behaviour. 2020;4:1208-1214. https://doi.org/10.1038/s41562-020-0912-z
6. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383-388. https://doi.org/10.1097/EDE.0b013e3181d61eeb
7. Brown JP, Hunnicutt JN, Ali MS, et al. Quantifying possible bias in clinical and epidemiological studies with quantitative bias analysis: common approaches and limitations. BMJ. 2024;385:e076365. https://doi.org/10.1136/bmj-2023-076365
8. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18
9. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Computational Biology. 2013;9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285
10. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proceedings of the National Academy of Sciences. 2018;115(11):2600-2606. https://doi.org/10.1073/pnas.1708274114
11. Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. Nature Human Behaviour. 2017;1:0021. https://doi.org/10.1038/s41562-016-0021
12. National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press; 2019. https://doi.org/10.17226/25303
13. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology; 2023. https://doi.org/10.6028/NIST.AI.100-1
14. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378