验证前工作论文

迈向客观且可审计的数据分析

GHEAS 可审计证据框架：设计可信度、测量可靠性与稳健性

Everion Cao

GHEAS，加拿大安大略省滑铁卢

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

方法学蓝图 v1.0 | 2026 年 7 月 21 日

摘要

背景

一项统计分析即使计算过程没有错误，也可能因为研究问题定义不清、数据来源不完整、测量不可靠、存在多种合理分析选择，或者因果表述超出研究设计能力，而形成脆弱甚至误导性的结论。选择性报告只是问题的一部分。可信的分析流程还必须展示数据构造、模型设定、时间对齐、测量误差、缺失和反证结果如何影响最终判断。

目的

提出一套可以适配不同领域的方法与软件契约，使统计结论更透明、更少受到立场影响、更容易复现，并使结论强度与实际证据相匹配。

框架

GHEAS 可审计证据框架（GHEAS Auditable Evidence Framework，GAEF）把分析视为一个有版本记录的证据过程，而不是选择一个偏好的模型。框架包含九个相互连接的部分：（1）主张与估计目标契约；（2）数据来源与测量契约；（3）研究设计与变量角色图；（4）相互分离的关联轨道与识别轨道；（5）预先规定的合理分析集合；（6）反证检验与失败门槛；（7）偏倚和不确定性的定量分析；（8）带有结论许可规则的证据综合；以及（9）具有明确 AI 边界的可执行审计记录。框架承认“证据不足”和“研究设计失败”是有效输出，并禁止统计显著性自动升级为因果结论。

验证方案

GAEF 将依次通过软件单元测试、已知真值模拟、对抗性数据扰动、公开基准数据、 PM2.5 与健康真实案例，以及至少一个跨领域案例进行检验。主要评价指标包括估计目标恢复偏差、区间覆盖率、错误结论率、错误检测能力、稳健性包络稳定性、计算复现性和审计完整度。

结论

本文是一份方法规范，而不是 GAEF 优于现有方法的证据。它提出一个可以被否定的命题：记录数据来源、展示合理替代分析、要求反证并限制结论强度的流程，应当减少没有充分依据的结论，同时不掩盖真正的不确定性。只有后续验证才能判断这一命题是否成立。

关键词： 数据分析；可复现性；稳健性；因果推断；测量误差；多宇宙分析； 人工智能；审计记录

1. 引言

统计研究最容易被看到的产物通常是一个系数、区间、预测或图形。但是，结论实际上来自一条更长的决策链：提出了什么问题、目标人群和对比是什么、原始观测怎样产生、哪些记录被保留、哪些变量被视为混杂因素或后果、时间如何对齐、为何选择某个模型，以及哪些证据有资格反对分析者偏好的解释。

即使不存在学术不端，传统报告也可能隐藏这条决策链。分析者在数据构造和模型选择中经常面对多个可以辩护的方案。只报告一个分析无法说明结论是否依赖这些选择。多宇宙分析和规格曲线分析正是为了展示这种依赖而提出的 [4,5]。同样，较小的 p 值不能表示效应大小、实际重要性或科学主张为真的概率 [1]。更复杂的模型或人工智能系统并不能自动修复这些问题。

不同方法学传统解决了决策链的不同部分。估计目标框架要求先定义要估计的量，再选择 估计方法 [2]。目标试验思想要求对齐纳入条件、暴露或处理分配、时间零点、随访、 结局和分析 [3]。负对照可以发现调整以后仍然存在的非因果关联 [6]。定量偏倚分析 可以评估测量、混杂和选择误差可能怎样改变结果，而不只是把它们写进局限性部分 [7]。 FAIR 数据原则和可复现计算规范则强调持久标识、机器可读元数据、完整转换记录和中间 结果保存 [8,9]。

这些方法都很重要，但常被当作互相分离的建议实施。GAEF 试图把它们连接为一个可执行 契约。它的目的不是创造一个适合所有问题的统计模型，而是约束问题、数据、设计、分析、失败检验、不确定性和结论语言之间的关系。

该框架来源于 PM2.5 污染与群体健康的开发经验。在这些工作中，一些看起来可以解释的 关联在加入国家特定趋势和时间安慰剂检验后发生变化；三个候选政策设计则由于污染物 第一阶段、事前趋势或趋势稳健性没有达到预定门槛，在估计健康结局以前就被停止。 这些经验只是提出框架的动机，不是对 GAEF 的验证。复杂观察性案例中的表现只有在方法 先通过已知真值和已知错误测试以后才具有解释意义。

2. 目标与适用范围

GAEF 面向使用观察、实验、行政、环境、实验室或连接数据的定量分析。它可以支持描述、关联、预测或因果问题，但不同主张需要满足不同条件。第一版实现以群体健康和环境数据 为重点，核心契约则希望能够迁移到其他领域。

框架有三个目标：

1. 减少没有记录的分析自由度；
2. 发现数值结果不能支持预定主张的情形；
3. 让从源数据到结论措辞的完整路径可以被独立审计。

GAEF 不保证真理，不取消专业知识，也不要求所有合理分析者得出相同结论。它不能恢复 从未测量的信息，不能在缺乏可信变异时识别效应，也不能证明不存在未测偏倚。它的作用 是让这些边界变得明确，并真实地限制结论。

3. 操作性定义

3.1 准确性

准确性是与已经定义的目标接近，而不是与已发表答案一致。数据生成过程已知时，可以用 偏差、均方根误差、区间覆盖率、校准和错误率来评价。在真实观察数据中，真值通常未知，此时框架报告识别假设、测量效度、稳健性和不确定性，不能把“多个结果一致”改称为真理。

3.2 客观性

客观性被视为流程属性。重要选择越明确，规则越能在后果未知时应用，合理替代方案越能得到对称检验，其他分析者越能重建结论如何获得许可，流程就越客观。预先登记有助于 区分事前预测与事后解释，但登记本身不能保证研究设计合理或分析正确 [10]。

3.3 不偏不倚

不偏不倚意味着支持和反对某一主张的证据采用相同的纳入、诊断和报告规则。零结果不会 自动优先于非零结果，统计显著也不能免除模型接受反证检验。失败的设计和相互矛盾的模型 必须留在记录中。

3.4 可审计性

可审计性意味着独立复核者能够把每个关键数字和主张追溯到有版本的输入、转换、模型、诊断、判断规则和软件环境。可复现是必要条件，但不是充分条件：一个完全可复现的分析仍可能回答了错误的问题。

3.5 校准后的结论

结论措辞不超过研究设计和诊断所支持的最强主张类别时，结论才是经过校准的。“数据中观察到”“变量之间存在关联”“模型能够预测”以及“在限定假设下支持因果解释”是不同 等级的输出。

4. GHEAS 可审计证据框架

4.1 组成一：主张与估计目标契约

每个项目首先建立机器可读契约，规定：

- 科学问题和主张类别；
- 目标人群、分析单位、暴露或干预、比较、结局和时间范围；
- 群体层面的汇总量和权重；
- 理论目标、经验估计目标、估计量，以及连接它们的条件；
- 允许使用预定结论语言的最低证据；
- 使当前数据无法识别目标的停止条件。

估计目标必须先于偏好估计器确定。不同权重、暴露增量、时间范围或结局定义代表不同估计目标，而不是可以随意互换的敏感性分析。契约可以修改，但必须记录日期、理由，并标明修改时相关结果是否已经可见。

4.2 组成二：数据来源与测量契约

每项输入必须记录提供者、版本或访问日期、许可、地理与时间覆盖、单位、测量类型、连接方法和已知处理历史。文件使用密码学哈希固定身份。原始文件保持不变，衍生文件只能由代码产生。

测量契约区分直接观察、仪器读数、行政记录、模型估计、插补和混合测量，并在适用时记录校准、空白、重复、批次、实验室、评审者一致性、连接和验证信息。原始数据不公开只能被记录为独立核验受限，不能被当作数据有错误的证明。

单位一致性是正式门槛。例如，PM10 不能直接改名为 PM2.5。换算必须具有明确的测量模型、转换参数的证据来源，并传播换算带来的不确定性。

4.3 组成三：研究设计与变量角色图

模型拟合以前，变量必须被赋予科学角色：暴露、结局、混杂、中介、碰撞变量或后果、精度变量、负对照、工具变量、效应修饰因素或描述变量。测量时间是角色的一部分。暴露以后测量或受结局影响的变量，不能在没有明确因果理由时被当作普通基线混杂因素。

研究设计图需要记录进入分析样本的选择机制、测量过程、干扰、缺失、时间顺序和识别变异来源。可以使用有向无环图辅助，但图形不能替代文字假设。

4.4 组成四：分离的证据轨道

GAEF 按能够支持的主张分开分析。

轨道 A：强化观察性分析。 轨道 A 估计描述性或条件关联量。最低要求包括未调整估计、预定调整集、适当的空间或分析单位及时间结构、替代暴露窗口、缺失分析、合理趋势控制和不确定性估计。轨道 A 可以说明关联稳定或脆弱，但多个模型一致也不能自动升级为因果结论。

轨道 B：明确识别分析。 轨道 B 要求随机分配机制或有文件依据的准实验变异来源，例如阈值、政策时间、边界、工具变量或自然事件。分析必须定义反事实对比，并检验设计特有的推论，包括暴露第一阶段、事前趋势或连续性、提前反应、溢出、处理状态反转和负对照。如果识别门槛失败，健康效应估计应当有意保持为空。

两个轨道不机械平均。一致可以提高证据的协调性；不一致则触发对人群、时间、测量、估计目标和假设的检查。

4.5 组成五：合理分析集合

在解释主要结果以前，项目需要定义同时具有科学合理性和实际影响的分析集合。维度可以包括样本资格、变量构造、暴露窗口、缺失处理、调整集、函数形式、权重、依赖结构和不确定性估计方法。

目的不是运行一切数学上可能的模型。每个分支必须有理由，逻辑上无效的分支应在看到结果以前排除。输出包括完整规格表、估计分布、方向与区间稳定性、最具影响的选择，以及通过全部诊断门槛的模型子集。这是在多宇宙和规格曲线思想基础上，把每个模型继续连接到结论许可，而不是把模型出现频率当作真理投票 [4,5]。

4.6 组成六：反证检验与失败门槛

GAEF 要求实施能够削弱偏好解释的分析。根据研究设计，可以包括：

- 未来暴露和不可能时间的安慰剂检验；
- 负对照结局或暴露 [6]；
- 事前趋势、提前反应、连续性、平衡或操纵检验；
- 合成零效应数据和结局置换；
- 单位、时间、编码和样本流程一致性检查；
- 预测问题中的样本外或时间验证；
- 逐单位、逐聚类排除或影响分析；
- 使用独立实现的估计器复现基准结果。

失败不等于证明真实效应不存在，而表示被检验的设计或模型不能许可预定主张。每个门槛对应的动作——继续、降级、隔离或停止——需要提前规定。

4.7 组成七：偏倚与不确定性分析

抽样误差只是不确定性的一部分。GAEF 分别处理：

- 随机误差；
- 测量误差与错误分类；

- 缺失与选择；
- 未控制或测量不良的混杂；
- 依赖和聚类；
- 模型与函数形式不确定性；
- 数据版本和连接不确定性；
- 可迁移性不确定性。

信息允许时，使用确定性或概率性定量偏倚分析，把合理误差参数传播到估计结果中 [7]。一种测量误差校正不能仅因它使效应变大而被接受；必须展示假设、验证来源以及对参数设定错误的敏感性。

4.8 组成八：证据综合与结论许可

框架输出的是证据账本，而不是一个享有特权的 p 值。每项主张都记录支持证据、反对证据、未解决假设、诊断失败，以及结论对分析选择的依赖。

参考证据阶梯为：

1. E0——无法分级： 数据、测量或设计不足；
2. E1——描述性： 在明确数据集中记录了某种模式；
3. E2——关联性： 关联在预定合理分析中持续存在；
4. E3——稳健关联或预测： 诊断、验证和偏倚分析支持稳定的非因果或预测主张；
5. E4——限定因果： 明确识别设计通过预定门槛，在陈述假设下支持因果解释。

这不是单一质量评分。强 E2 关联和没有结论的 E4 尝试回答的是不同问题。AI 没有权限提高证据等级。

4.9 组成九：可执行审计记录与受限 AI

每次运行生成清单，包含协议哈希、输入哈希、软件版本、随机种子、样本流程、模型清单、诊断、排除、修订和产物。按照可复现计算必须完整记录结果产生过程的原则，中间结果也 需要保存 [9]。在条件允许时，数据和元数据应当可发现、按适当权限访问、可互操作和 可复用 [8]。

AI 可以协助代码生成、数据质量检查、文献分流、模型运行、比较和报告起草。它不能静默改变问题、删除矛盾模型、编造缺失来源、在没有复核时裁决高风险排除，或者升级因果语言。AI 输出必须能够追溯到模型和会话或工作流版本，并接受确定性测试和人的最终责任。这与 AI 系统应在完整生命周期中接受有效性、可靠性、问责和透明度评价的广泛 要求一致 [13]。

5. 参考软件架构

框架映射为七个软件层：

1. 协议层： JSON 模式和人类可读研究契约；
2. 数据登记层： 来源、许可、测量、覆盖和哈希；
3. 验证层： 模式、单位、范围、主键、时间、缺失和重复检查；
4. 分析引擎： 面向不同主张的轨道 A、轨道 B 及合理分析集合；

5. 反证引擎： 负对照、安慰剂、设计门槛、影响分析和对抗性检查；
6. 证据协调器： 结构化比较和有规则约束的结论等级；
7. 报告与审计生成器： 完整表格、图形、局限、清单和机器可读结果。

系统必须“关闭式失败”：缺少必要输入、假设、诊断或来源字段时，不能静默继续生成更强结论。警告必须改变分析状态，而不能只出现在日志中。

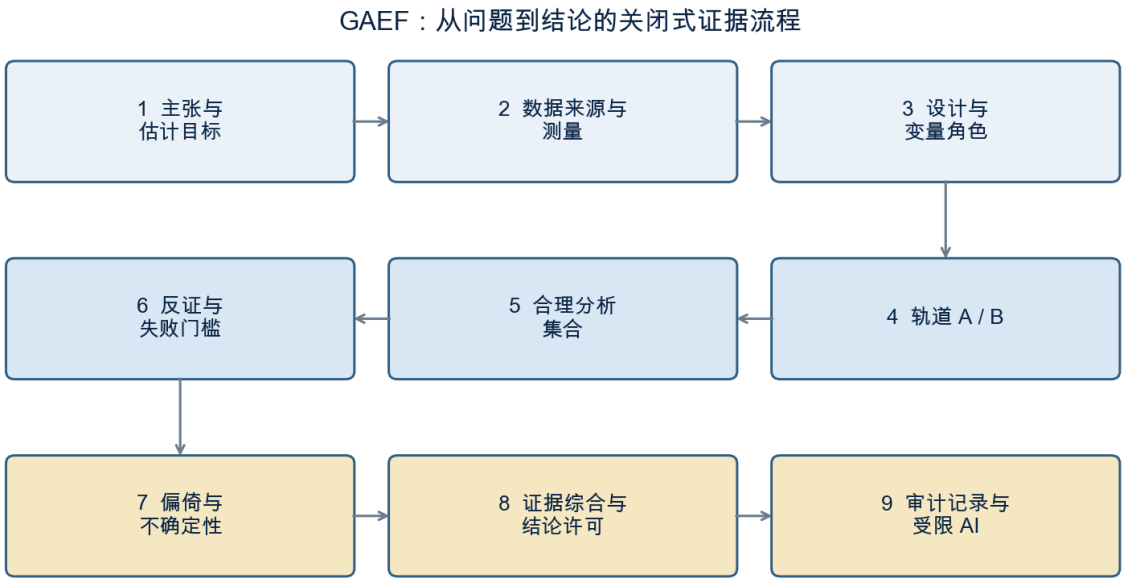

图 1．GAEF 以关闭式流程连接研究规范、分析、反证和结论许可。

6. 验证计划

6.1 V1：软件正确性

单元测试验证转换、估计器、标准误、置信区间、哈希和判断规则。确定性数据夹具检验已知零、正、负和异质效应。独立实现结果需要在数值容差内一致。格式错误的输入必须触发预定失败状态。

6.2 V2：已知真值模拟

因子模拟将改变样本量、效应大小、非线性、异质性、聚类、序列相关、暴露持续性、混杂强度、干扰和政策采纳，并包含零效应和方向反转。主要指标为偏差、均方根误差、 95% 区间覆盖率、方向错误、效应夸大和错误升级结论等级的频率。

模拟代码和评价标准必须在查看总体性能以前固定。方法不能仅因在专门为其设计的数据生成过程中表现良好就被判定通过。

6.3 V3：对抗性扰动

在干净数据中主动加入已知错误：单位换算错误、重复记录、时间戳偏移、实验室批次效应、暴露测量误差、结局泄漏、非随机缺失选择、编码改变、极端高影响点和选择性删除模型。评价错误检测灵敏度、误报率、严重程度判断以及是否正确降低了结论等级。

6.4 V4：公开基准

重新分析具有公开数据和已记录分析或已知设计属性的数据集。与原论文一致只被视为复算结果，不能证明原论文正确。条件允许时，在锁定协议和代码以前，最终评价者不知道基准 结果。

6.5 V5：PM2.5 与健康案例

现有 GHEAS 资产提供三类互补检验：

- 国家—年份 PM2.5 与预期寿命面板，用于模型依赖和时间反证；
- 已审计的政策或自然实验候选，用于轨道 B 失败门槛；
- 长期 PM2.5 与死亡证据库，用于研究重叠、暴露测量、偏倚风险和证据综合。

框架实现冻结后，比较传统分析、轨道 A、轨道 B 和最终证据账本。不能一边用本案例调整规则，一边又把同一案例当成独立验证。

6.6 V6：跨领域迁移

在宣称具有通用性以前，至少选择一个空气污染以外的案例。候选包括饮水污染、气候与健康、营养、实验室测量或其他纵向公共数据。允许领域专用模块，但核心主张和审计契约必须保持不变。

7. 评价与验收标准

在验证结果产生以前，本文预先规定以下评价指标类别：

- 数值有效性： 已知真值下的估计偏差、误差、区间覆盖和校准；
- 判断有效性： 包括零效应和混杂场景下不当因果或方向性结论的频率；
- 错误发现： 对注入的数据和流程错误的灵敏度、特异度和严重程度分类；
- 稳健性透明度： 合理规格展示比例、导致实质变化的选择及矛盾结果完整度；
- 可复现性： 相同输入和环境产生相同清单与结果，独立实现保持在规定容差内；
- 审计完整度： 关键主张能够追溯到来源、转换、模型、诊断和规则的比例；
- 可用性： 所需时间、专业能力、人工干预，以及正确但无法行动的警告频率。

具体通过阈值将在最终测试以前写入验证协议，避免在知道表现以后选择有利标准。

8. 与现有方法的关系

GAEF 不替代领域专用统计方法或报告规范，而是连接已有思想：

- 明确估计目标以及设计、分析和解释之间的对齐 [2,3]；
- 多宇宙和规格曲线分析 [4,5]；
- 负对照和定量偏倚分析 [6,7]；
- FAIR 数据管理和可复现计算 [8,9]；
- 区分预先登记与探索性工作 [10]；
- 关于可复现和可重复研究的机构性建议 [11,12]；
- AI 辅助系统的全生命周期问责 [13]。

本文提出的贡献是带结论门槛的可执行整合，而不是发明上述组成方法。领域报告规范仍然必要。例如，预测模型研究仍应采用 TRIPOD+AI 等适用的透明报告规范 [14]。

9. 治理与人的责任

“AI 做分析”可能掩盖责任。GAEF 规定人仍然负责科学问题、伦理边界、变量角色、因果假设、访问权限、解释和公开。数据限制能够观察的内容。AI 和软件负责执行已声明规则、发现不一致、比较替代方案和记录过程。

一切重要人工覆盖都需要理由和身份。人工修改并不被禁止，未记录修改才被禁止。复核者应当能够对门槛判断提出异议，但原判断和覆盖后的判断都必须可见。

隐私与公平需要领域专用评价。一个可审计模型仍可能造成伤害、不具代表性或被用于目标人群以外。GAEF 因此区分统计上的不偏不倚与社会公平，不宣称前者自动保证后者。

10. 局限与可能的失败方式

框架可能造成过度复杂、错误安全感，或产生用户无法解释的大量诊断。预先规定可能冻结错误假设。合理分析集合仍受设计者对“合理”的判断影响。失败检验可能把握度不足；负对照通过不能证明不存在混杂；定量偏倚分析依赖不确定的偏倚参数；可复现代码也可以复现系统性错误。

自动化还会带来其他风险：错误代码可能被规模化运行；AI 解释可能比证据更确定；软件默认值可能成为隐藏的方法权威；用户可能把证据等级当作判断的替代品。因此，框架本身必须接受对抗性检验，并报告自身错误。

初始开发团队和案例范围有限。在获得独立使用者、跨领域案例和独立实现以前，不能提出广泛的有效性或通用性主张。

11. 讨论

GAEF 最核心的设计选择，是让结论成为一个证据过程的输出，而不是对单一模型的语言解释。该过程可以输出数值、限定主张、降级主张或停止。把停止视为合法结果，会改变研究激励：一项分析即使没有产生健康效应估计，也可以具有科学价值。

这也改变了论文的作用。本文是用来检验软件和验证结果的规范，观察到失败以后应当修改。后续论文可以报告成功和失败的验证结果，但是否发表不是验收标准。真正标准是：方法在已知真值下是否达到预定表现，能否发现有意义的错误，是否保留相反证据，以及能否在真实应用中正确校准结论。

PM2.5 案例的价值恰恰在于它很困难：暴露主要来自模型，不同研究的人群和时间尺度不同，混杂严重，政策可能只有很弱的第一阶段，既有结果也存在异质性。因此，它适合作为可用性和可审计性的压力测试，却不适合作为统计准确性的唯一检验，因为真实群体效应未知。

12. 结论

GAEF 提出一条以方法为先的路径，使数据分析更加客观和可审计。它不要求某一个模型、 分析者或 AI 系统天然无偏，而是通过明确目标、记录来源、评价测量、比较合理替代方案、反证、定量不确定性、结论门槛和完整审计记录约束整个过程。

目前这一框架仍是一个假设。下一步不是宣传它，而是用已知真值模拟、对抗性错误、公开基准和困难真实数据尽力推翻它。如果方法不能通过这些检验，无论论文写得多么有说服力，都应当被修改或放弃。

声明

经费： 本研究没有外部经费支持。

利益冲突： 作者声明不存在利益冲突。

数据和代码： 工作框架、协议、代码、测试和案例资产保存在带版本记录的 GHEAS 项目中。公开发布将在验证完成并经过隐私和许可检查后进行。

AI 协助： OpenAI Codex 在项目中称为 Jace，在人的指导下协助文献整理、写作、代码开发和文档制作。它不是作者，也不承担研究责任。所有科学主张和发布决定均由署名作者负责。

参考文献

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