arXiv:2603.03445v1 [stat.ME] 3 Mar 2026

确定性界限：科学可靠性的结构极限

Marco Pollanen 特伦特大学数学与统计系 加拿大安大略省彼得伯勒 K9L 0G2

作者说明

ORCID：Marco Pollanen，https://orcid.org/0000-0001-5356-1889

通信请联系：marcopollanen@trentu.ca

投稿至 *Meta-Psychology*。如欲参与开放同行评审，请发送电子邮件至 open.peer.reviewer@gmail.com。点击以下链接即可查看 *Meta-Psychology* 所有审稿中文章的完整编辑流程：https://tinyurl.com/mp-submissions。搜索作者姓名即可找到这篇预印本。

确定性界限：科学可靠性的结构极限

对可重复性危机的解释往往强调不端行为、有问题的研究实践或激励错配，并暗示仅靠行为改革就足够了。本文指出，这场危机的一个重要组成部分源于架构：在基于二元显著性的发表架构中，即使研究人员完全尽职尽责，他们所能实现的可靠性也存在结构极限。

一项发现的后验对数优势等于先验对数优势加上 $\log \Lambda$，其中$\Lambda = (1-\beta)/\alpha$ 是实验杠杆。从架构角度解读，这意味着一项硬性约束：一旦证据被简化为二元显著性决策，决策规则对后验对数优势的贡献恰好为 $\log \Lambda$ 。当且仅当 $\pi \geq \pi_{\text{crit}}$时，目标可靠性 τ 才是可行的；在固定 α 的情况下，通常不能仅靠增大样本量来满足这一条件。两种不同机制都可能使有效杠杆降至 1：观察性研究中的持续性未测量混杂，以及发表压力下无界的设定搜索；这两种情况都不要求存在恶意。这些结果适用于基于二元显著性的决策架构，并不限制基于完整似然或更丰富的连续证据摘要进行推断。两项坍缩结果将这些机制形式化，而复制流程定理和最小管线深度推论则确定了实现突破所需的定量证据标准。

将框架应用于改革前心理学经过独立记录的参数（$\pi \approx 0.10$，检验效能 ≈ 0.35），可得重复率为 36%，与开放科学协作组织公布的数据一致。该框架还为科学哲学提供了定量衔接，包括波普尔式证伪、库恩式范式转换和拉卡托斯式退化研究纲领。在先验概率较低、低于单项研究可行性阈值的情境下，自然的证据单位是复现流水线，而非单项实验。

关键词：：可重复性危机、可重复性危机、元科学、发表偏倚、统计显著性、假发现率、阳性预测值、多重检验

1 引言

1.1 一个谜团及其解答

2015年，开放科学协作组织报告了心理学领域100次重复研究尝试的结果。根据基于显著性的重复研究判据，36%的重复研究尝试得到了方向相同的显著结果（开放科学协作组织，2015年）。早期改革讨论大多从行为层面解释这一结果：研究者被认为会进行P值操纵、选择性报告，并对奖励新颖性而非严谨性的激励机制作出反应。因此，隐含的补救办法也着眼于行为：培养更优秀的科学家、办好期刊，并建立更好的规范。

本文提出了另一种解释：它或许更令人不安，但也更便于分析。36%这一数字不必主要被视为不当行为的表征。它与一种结构性结果相符，而这种结果是由数学恒等式在经独立记录的运行条件下所蕴含的。从这一视角看，核心问题不在于个别研究者的道德品格，而在于证据架构的配置。

统计显著发现的可靠性受一个核算恒等式支配。令 π 表示待检验假设为真的先验概率，令$\Lambda = (1 - \beta)/\alpha$ 表示实验杠杆，即检验效能与假阳性率之比。则阳性预测值满足 $\text{PPV} = \pi\Lambda/[\pi\Lambda + (1 - \pi)],$，而一次检验对真实效应的后验对数优势恰好贡献 $\log \Lambda$ 。这是贝叶斯定理的一种常见形式。本文的贡献在于架构层面，而非代数层面：一旦证据被简化为二元显著性决策，实验对后验对数优势的贡献就恰好是 $\log \Lambda$ 。在这种架构下，同一决策规则无法提取额外的证据增益。如果 π 较小，或 Λ有限，那么无论样本量多大、研究者多么勤勉，后验概率仍然偏低。

元科学文献中记录的参数范围在重复研究项目之前（$\pi \approx 0.10$，作为与Ioannidis（2005）中的情景一致的校准参数处理；中位数检验效能为 ≈ 0.35，Cohen（1962）；Sedlmeier 与 Gigerenzer（1989））意味着阳性预测值为 ≈ 0.44 ，重复率约为 36%。

这些参数并非经过校准以匹配开放科学协作组织的发现；这种一致性是将该框架应用于独立记录的条件后所产生的结构性推论。

1.2 架构论证

从根本上说，本文将解释从行为层面转向架构层面。行为解释将问题归因于研究者的选择：P值操纵、选择性报告以及对激励措施作出反应的行为。架构解释则将问题归因于研究系统本身的结构。在本文中，“架构层面”指由领域层面的运行参数（π、 α、$1-\beta$）和发表决策规则所决定，而非由研究者各自的行为习惯所决定。这一区别很重要，因为两种解释提出的补救措施不同，所蕴含的道德意义也不同。

架构解释并不意味着不当行为可以免责。有问题的研究实践确实存在，而规格搜索崩溃（第 5 节）对其造成的损害进行了形式化描述。但架构视角明确揭示了行为解释可能掩盖的一点：即便假设某个领域中的研究者全都诚实、严谨且技术精湛，在这些运行参数下，也无法在 $\pi = 0.10$ 和 $\alpha = 0.05$时将PPV提高到69%以上。

Ioannidis（2005）表明，在现实假设下，阳性预测值较低。本文表明，在固定的 α 以及基于二元显著性的发表架构下，阳性预测值可能受到结构性限制，远低于常见的可靠性目标值；这是该架构本身的约束，而非偶然的缺陷。Ioannidis诊断的是一个快照（给定参数配置下的阳性预测值），而本文的框架诊断的是一条轨迹：代际动态、领域寿命和退化性研究纲领判据刻画了随着领域成熟和后续研究不断累积，可靠性如何演变。确定性界限所依据的阳性预测值恒等式，已见于早期关于假阳性率的研究（例如，参见Wacholder等，2004）；本文的贡献在于从设计层面进行阐释、设定阈值、分析坍缩，并探讨其制度层面的影响。

有必要澄清范围。该框架讨论的是

二元显著性主张的可靠性：即一项统计显著发现反映真实效应的后验概率。连续估计、贝叶斯推断以及更丰富的基于似然的方法可以从数据中提取更多信息。本分析刻画了二元发表架构所施加的上限；这些约束源于发表架构，而非统计理论本身。

本文作出四项贡献。第一，它证明了一项结构性不可行性结果：在临界先验 π_{crit} 以下无法实现高可靠性，并识别出两条机制上截然不同的坍缩路径（观察性混杂与设定搜索），二者共享不变量 $\Lambda_{\text{eff}} \to 1$。第二，它借助詹森不等式推导出异质性代价，并建立一条将本框架与观测到的复现率联系起来的 Replication Bridge。第三，它发展了代际动态，从而给出定量的退化性研究纲领判据。第四，它证明了复制流程定理，确立复现流水线是低先验情境下的自然证据单位。

自始至终，其目的并非取代现有的元科学工具，而是提供一个结构性判据，用以判断高可靠性在数学上究竟何时首先是可行的。

确定性界限诊断工具、流水线计算器与可靠性景观的交互式实现可在 https://mpollanen.github.io/certainty-bound-tool/ 获取。[1]

1.3 路线图

第2节构建形式化框架，第3节证明多数为假定理和发现成本定理，第4节确立 Replication Bridge，并说明其与开放科学协作组织发现的一致性。第5至第6节分析崩溃机制和逃逸路径，包括复制流程定理。第7至第9节阐述领域动力学、可靠性景观以及可靠研究的设计要求；第10节和第11节探讨其影响及其与科学哲学的联系。

[1] 该工具亦存档于开放科学框架：https://osf.io/c5wun。

2 确定性界限：形式化框架

2.1 设置

定义1（状态空间与先验）。 设 $H \in \{0,1\}$ 表示某个假设的真值，其中 $H=1$ 表示效应存在， $H=0$ 表示效应不存在。设$\pi := \mathbb{P}(H=1)$ 表示为调查而选定的假设为真的先验概率 。

先验 π 并非所有可设想假设中为真的假设所占的比例；它是指，在促使某个假设被选中进行检验的理论推理、初步数据和激励结构的背景下，该假设反映真实效应的概率。具有II期证据的临床试验可能有 $\pi \approx 0.30$；全基因组扫描可能有 $\pi \approx 10^{-5}$；探索性社会心理学实验则可能有 $\pi \approx 0.05$–0.15。因此，先验是一个领域的假设生成过程的属性，概括了其研究议程有多大胆或多保守。

定义 2（检验结果与错误率）。 统计检验会产生二元结果：显著（$T=1$）或不显著（$T=0$）。检验由显著性水平 $\alpha := \mathbb{P}(T=1 \mid H=0)$ 和检验效能 $1-\beta := \mathbb{P}(T=1 \mid H=1)$ 表征。

注释 3（名义水平与有效错误率）。 我们区分名义 错误率与有效 错误率。由于分析灵活性或选择性报告，有效 α_{eff} 可能高于名义 α 。当我们不带下标写作 α 和 $1-\beta$ 时，指的是决定实际接受决策 T的错误率；不存在设定搜索或选择性报告时，这些错误率与名义错误率一致。除非明确标注（例如， α_{eff}、 Λ_{eff}），否则 α 和 β 指名义错误率。第 2–4 节讨论名义参数；第 5 节通过设定搜索和混杂引入有效错误率。阳性预测值恒等式（1）适用于决定实际决策过程的任意错误率。

模型假设。 全文中我们假设 $\pi \in (0,1)$、 $\alpha \in (0,1)$、$\beta \in [0,1)$，以及 $\Lambda = (1-\beta)/\alpha \in (0,\infty)$。所有概率陈述都涉及用于生成二元主张 T 的决策规则；也就是说， α 和 $1-\beta$ 是实际检验研究流程的运行

特征。每当对 Λ 取乘积或幂时，都会显式说明条件独立假设。

定义 4（实验杠杆）。 统计检验的实验杠杆 为

$$\Lambda := \frac{1-\beta}{\alpha} = \frac{\text{power}}{\text{false-positive rate}}.$$

杠杆衡量假设为真时出现显著结果的可能性是假设为假时的多少倍。检验效能为 0.80 的检验且 $\alpha = 0.05$ 具有 $\Lambda = 16$。

备注 5（最小判别力与符号反转）。 阈值 $\Lambda = 1$（等价地， $1 - \beta = \alpha$）是显著结果的证据与反证据之间的界限。由（1）可知，若 $\Lambda = 1$ ，则阳性预测值 $= \pi$，因此显著结果并未提供先验之外的信息。若 $\Lambda < 1$，则阳性预测值 $< \pi$，因此相对于先验，显著性属于反证据。在这种状态下，框架仍然有效，但证据更新的方向会反转。因此，以下许多阈值结果都假设满足最小判别力，即 $\Lambda > 1$。

定义 6（阳性预测值）。 阳性预测值：$= \mathbb{P}(H = 1 \mid T = 1)$。

2.2 确定性界限

定理7（确定性界限）。 对于任何杠杆为 Λ 、应用于先验为 $\pi \in$ （0，1）的假设的统计检验：

$$\text{PPV} = \frac{\pi\Lambda}{\pi\Lambda + (1-\pi)}. \tag{1}$$

等价地，用对数优势形式表示：

$$\log \frac{\text{PPV}}{1-\text{PPV}} = \log \frac{\pi}{1-\pi} + \log \Lambda. \tag{2}$$

达到 阳性预测值 $\geq \tau$ 需要

$$\Lambda \geq \Lambda_{\text{req}}(\tau, \pi) := \frac{\tau}{1-\tau} \cdot \frac{1-\pi}{\pi}. \tag{3}$$

证明。 根据贝叶斯定理：阳性预测值 $= (1-\beta)\pi / [(1-\beta)\pi + \alpha(1-\pi)] = \pi\Lambda / [\pi\Lambda + (1-\pi)]$。取对数优势比可得式（2）。令阳性预测值 $\geq \tau$ 并求解，可得式（3）。 □

对数优势形式使信息含量一目了然：一旦证据被简化为显著性决策，检验对后验对数优势的贡献恰好是 $\log \Lambda$ ，除此之外别无其他。统计显著性并非证据本身的属性，而是决策规则的属性。在这一架构下，显著性决策的最大证据贡献由方案的杠杆作用决定。研究者的审慎程度之所以重要，是因为它会改变运行参数，包括有效错误率、检验效能、识别或受检假设的先验概率；但对于给定的二元决策，它无法提取超出 $\log \Lambda$ 所能提供的信息。

确定性界限对于二元决策是精确的（$T \in \{0, 1\}$）。它并不限制使用更丰富的证据摘要（完整似然、后验分布或连续效应量估计）所能达到的水平，因为这些摘要能够从相同数据中提取更多信息。

该框架将三个量联系起来，但并未将它们等同：在二元发表筛选下显著发现的阳性预测值（可靠性）、特定重复研究设计下的重复率（第4节），以及受检假设的潜在先验概率 π 。混淆这些量是常见的困惑来源；它们在分析中各自发挥不同作用。

Exam例 8 （所需杠杆作用）。 对于目标值 $\tau = 0.95$，优势比 $\tau/(1-\tau) = 19$:

先验 π	先验优势 $(1-\pi)/\pi$	所需 Λ_{re} q
0.50	1	19
0.10	9	171
0.01	99	1,881
10^{-5}	99,999	1.9×10^{6}

随着先验概率降低，所需杠杆作用按 $O(1/\pi)$ 的比例增长。

2.3 固定-α 上限

定理 9 （固定α上限）。 对于任何显著性水平固定为$\alpha \in$ （0，1）的统计检验，无论样本量多大：

$$\mathrm{PPV} \leq \mathrm{PPV}^{\mathrm{ceil}}(\pi, \alpha) := \frac{\pi}{\pi + \alpha(1-\pi)}. \tag{4}$$

此上限是紧的，当检验效能趋于 $\to 1$时可达到。

证明。 由于阳性预测值在 Λ （其中

— $\frac{\partial}{\partial\Lambda}$ 阳性预测值 $= \pi(1-\pi)/[\pi\Lambda + (1-\pi)]^2 > 0$），且 $\Lambda = (1-\beta)/\alpha \leq 1/\alpha$ 对于 $\beta \geq 0$，当 $\beta \to 0$ 时取等号：阳性预测值$^{\text{ceil}} = \pi/[\pi + \alpha(1-\pi)]$。 □

推论10 （ $\alpha = 0.05$ 处的上限值）。

先验 π	阳性预测值$^{\text{ceil}}$	解释
0.50	*95.2%*	勉强达到*95%*目标值
0.30	*89.6%*	无法达到*90%*阳性预测值
0.10	*68.9%*	阳性预测值不能超过 *69%*
0.05	*51.3%*	仅略高于半数为真
0.01	*16.8%*	假阳性占多数

对于改革前心理学（$\pi \approx 0.10$, $\alpha = 0.05$），无论样本量多大，可达到的阳性预测值最高为 69%，远低于解读时通常假定的 95% 可靠性。上限仅取决于 α 和 π；增加样本量无法改变这一上限。

2.4 临界先验概率与不可行性指数

定义11 （不可行性指数）。 不可行性指数 为

$$\Psi(\tau, \pi, \alpha, \beta) := \frac{\Lambda_{\text{req}}(\tau, \pi)}{\Lambda(\alpha, \beta)} = \frac{\tau}{1-\tau} \cdot \frac{1-\pi}{\pi} \cdot \frac{\alpha}{1-\beta}. \tag{5}$$

定理 12 （临界先验与结构性不可行）。 对于给定的 $\alpha \in (0,1)$, $\beta \in [0,1)$, 且目标值 $\tau \in (0,1)$:

(a) 可行性： 当且仅当 $\pi \geq \pi_{\text{crit}}(\tau, \alpha, \beta)$ 时，目标阳性预测值可实现，其中 *re*

$$\pi_{\text{crit}}(\tau, \alpha, \beta) = \frac{\tau\alpha}{(1-\tau)(1-\beta) + \tau\alpha}. \tag{6}$$

(b) 不可行性： 如果 $\Psi > 1$ （等价地， $\pi < \pi_{\text{crit}}$），那么在给定运行参数下，目标阳性预测值在结构上无法达到。在固定 α 的情况下，仅增加样本量无法突破固定-α 上限；而任何不改变运行参数的研究实施改进也无法克服这一限制。

证明。（a）部分：根据定理7，当且仅当 $\Lambda \geq \Lambda_{\text{req}}(\tau, \pi)$ 时，PPV $\geq \tau$。将 $\Lambda = (1-\beta)/\alpha$ 代入并求解 π，得到（6）。

部分（b）：根据定义11，$\Psi > 1$ 当且仅当 $\Lambda_{\text{req}}(\tau, \pi) > \Lambda$；等价地（根据部分（a）），当且仅当 $\pi < \pi_{\text{crit}}(\tau, \alpha, \beta)$。在这种情况下，在所述运行参数下，阳性预测值 $< \tau$。此外，在 α 固定的条件下，定理 9 表明阳性预测值 $\leq \pi/[\pi + \alpha(1-\pi)]$,，因此，一旦目标值高于上限，单靠增加样本量就无法克服这一限制。任何不改变 (π, α, β) 的干预也不会改变 Λ，因而无法改变由此推得的阳性预测值。

□

该定理重新界定了问题：它不再询问阳性预测值是否足够高，而是询问领域先验是否超过 π_{crit}。分析焦点由单项研究质量转向科研体系的运行参数。

表1不同配置下、目标值为 $\tau = 0.95$ 时的临界先验 $\pi_{\text{crit}}(\tau, \alpha, \beta)$。

α	检验效能	π_{crit}	含义
0.05	0.80	54.3%	受检假设中多数必须为真
0.05	0.50	65.5%	三分之二必须为真
0.05	0.30	76.0%	四个中必须有三个为真
0.01	0.80	19.2%	五个中必须有一个为真
0.005	0.80	10.6%	九个中必须有一个为真
5×10^{-8}	0.80	1.2×10^{-6}	840 分之一 ,000

有三个规律值得注意。将 α 从 0.05 降至 0.005，会使 π_{crit} 从 54% 降至 11%，改善了五倍。降低检验效能会提高 π_{crit}：检验效能为 0.30 而非 0.80 时，要求四分之三的受检假设为真。全基因组关联研究阈值（$\alpha = 5 \times 10^{-8}$；参见 Pe’er 等，2008）会将 π_{crit} 降至约 84 万分之一，从而能够在先验概率极低的情况下可靠地发现结果。

2.5 异质性代价

命题 13 (先验异质性惩罚)。 设 Π 是定义在 $(0,1)$ 上的非退化随机变量，其均值为 $\bar{\pi}$，方差为 $\sigma_\pi^2 > 0$。对于固定的 $\Lambda > 1$:

$$\mathbb{E}[\mathrm{PPV}(\Pi)] < \mathrm{PPV}(\bar{\pi}), \tag{7}$$

差距近似为 $\Lambda(\Lambda-1)\sigma_\pi^2/[\bar{\pi}(\Lambda-1)+1]^3$。当$\Lambda > 1$ 且 Π 非退化时，严格不等式成立。

证明。 参见附录 11.1。 □

异质性会给可靠性带来代价。如果一个领域混合了先验概率不同的假设类别，例如将探索性试探与验证性后续研究混在一起，那么即使杠杆保持不变，其平均 阳性预测值也会低于由其平均 先验概率所隐含的阳性预测值。其机制是凹性，而非偏差。

这种惩罚是实实在在的：在 $\Lambda = 16$ 时，$\mathrm{PPV}(0.10) = 0.64$。但以相等概率混合 $\pi \in \{0.02, 0.18\}$，平均阳性预测值为 ≈ 0.51，尽管先验概率均值相同。凹性对低-π 观测的惩罚，大于对高-π 观测的奖励，因此，混合领域中低先验概率的一半对平均值的拖累，大于高先验概率的一半对平均值的提升。采用标准化方案和注册报告的领域可以降低这一代价；混合探索性与验证性研究的领域则要全额承担。

由此可得出一个可检验的推论：在杠杆大致相同的情况下，先验异质性更大的领域，其重复率应低于单一先验校准所预测的水平。

备注 14（聚合偏差）。 将“心理学”这类学科视为一个整体，并假设其只有一个先验 $\pi \approx 0.10$，是一种简化；实际上，各子领域之间差异很大。然而，由于阳性预测值函数严格凹（命题 13），这种聚合会引入乐观偏差。由高可靠性和低可靠性子领域混合组成的领域，其总体重复率会低于参数处于平均水平的同质领域。因此，这里提出的不可行性结果应解读为对整个领域可靠性的保守上界：结构性现实很可能比汇总模型所显示的情况更糟。

结构性现实很可能比汇总模型所显示的情况更糟。

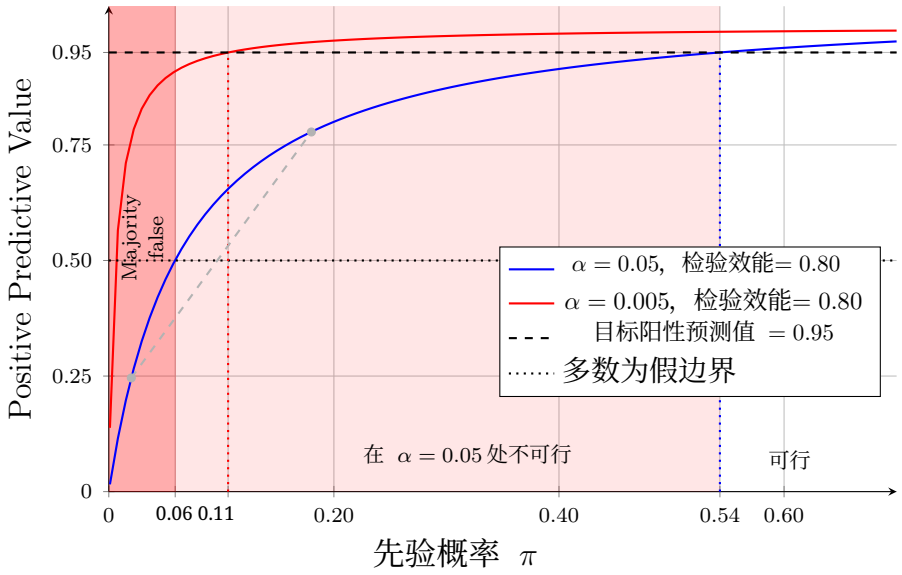


图1

两个显著性阈值下，阳性预测值与先验概率的关系。深色阴影区域（$\pi < 0.059$）是在常规参数下的多数为假状态。浅色阴影区域（$0.059 < \pi < 0.543$）对于目标值 $\tau = 0.95$ 在 $\alpha = 0.05$下不可行。将 α 收紧至 *0.005* 后，临界先验变为 *0.106*。连接 $\pi = 0.02$ 和$\pi = 0.18$ 的弦位于曲线下方，形象地展示了异质性代价。

注释 15（独立性）。 下文的若干结果（复制流程定理、规格搜索模型和发现成本定理）将研究或重复检验视为统计独立。这是一个理想化的基准。研究之间的依赖性（共享样本、相关假设或序贯更新）通常会削弱杠杆倍增，并增强不可行性结果在实践中的意义。

假设图。 为清晰起见，我们区分不同结果所使用的假设。确定性界限（定理7）和固定-α 上限（定理 9）仅要求二元粗化和所述错误率。复制流程定理（定理 33）和发现成本（定理 17）还假设研究在给定 H时条件独立。观察性崩塌（定理 21）

是在持续混杂条件下关于样本量的渐近结果。规格搜索崩溃（命题 28）采用规格尝试的序贯独立性假设。代际动态（定理37）和领域寿命（定理35）是采用固定 Λ和确定性传播的风格化动态模型。

3 两种不可能状态

3.1 多数为假定理

定理 16 （多数为假阈值）。 对于杠杆为 Λ 的检验，显著发现中多数为假阳性（阳性预测值 $< 1/2$）当且仅当

$$\pi < \pi_{1/2} := \frac{1}{1+\Lambda} = \frac{\alpha}{(1-\beta)+\alpha}. \tag{8}$$

当 $\alpha = 0.05$ 且检验效能为 0.80： $\pi_{1/2} = 1/17 \approx 5.9\%$。当 $\alpha = 0.05$ 且检验效能为 0.35：$\pi_{1/2} = 12.5\%$。

证明。 阳性预测值 $< 1/2$ 当且仅当 $\pi\Lambda < (1-\pi)$ 当且仅当 $\pi < 1/(1+\Lambda)$。 □

在探索性研究中，很容易越过多数为假阈值。如果某个领域中真正成立的假设不到大约八分之一到十七分之一（具体取决于检验效能），那么即使不存在任何有问题的研究实践，该领域的文献中显著发现也会以假阳性居多。

3.2 发现成本

定理 17 (发现成本**)**。 在恒定参数 (π, α, β) 且研究彼此独立的领域中，假阳性与真阳性的长期预期比率（等价地说，在 $N \to \infty$ 时， N 项研究的预期数量之比）为

$$W = \frac{(1-\pi)\alpha}{\pi(1-\beta)} = \frac{1-\mathrm{PPV}}{\mathrm{PPV}}. \tag{9}$$

证明。 参见附录11.2。 □

例 18 (跨领域的科学成本**)**。

领域	π	检验效能	α	PPV	W
候选基因	0.02	0.50	0.05	17%	4.9
改革前心理学	0.10	0.35	0.05	44%	1.3
营养流行病学	0.08	0.60	0.05	51%	1.0
统计功效充分的随机对照试验	0.30	0.80	0.05	87%	0.15
GWAS	10^{-5}	0.80	5×10^{-8}	99.4%	0.006

候选基因文献中，每个真阳性几乎对应 5 个假阳性，而每项真正的发现大约需要 100 项研究。改革前心理学中，每项真正的发现大约需要 29 项研究。

4 Replication Bridge

4.1 Replication Bridge 定理

定理 19 (ReplicationBridge)。 考虑一项阳性预测值为阳性预测值$_o$ 的原始研究，以及一项在显著性水平 α_r 下进行、检验效能为 $1-\beta_r$的独立重复。原始研究结果显著的条件下，重复研究结果显著的概率为：

$$\mathbb{P}(T_r = 1 \mid T_o = 1) = \mathrm{PPV}_o \cdot (1-\beta_r) + (1-\mathrm{PPV}_o) \cdot \alpha_r. \tag{10}$$

反过来，给定一个观测到的重复率 R，并且满足 $1-\beta_r > \alpha_r$：

$$\widehat{\mathrm{PPV}} = \frac{R-\alpha_r}{(1-\beta_r)-\alpha_r}.$$

证明。 根据全概率定律，在以 H 为条件，并利用在给定 H的条件下 T_r与 T_o 的独立性：

$$\begin{aligned}\mathbb{P}(T_r = 1 \mid T_o = 1) &= \mathbb{P}(H = 1 \mid T_o = 1)\,\mathbb{P}(T_r = 1 \mid H = 1) + \mathbb{P}(H = 0 \mid T_o = 1)\,\mathbb{P}(T_r = 1 \mid H = 0) \\ &= \mathrm{PPV}_o \cdot (1-\beta_r) + (1-\mathrm{PPV}_o) \cdot \alpha_r.\end{aligned}$$

解出PPV$_o$ 即可得到反演公式。 □

Replication Bridge假定，在以 H为条件时，原始研究与重复验证相互独立；共享的材料、实验人员或参与者群体可能违反这一假定，从而削弱实际实现的杠杆倍增。

4.2 与观察到的重复率的一致性

框 1：改革前心理学的参数校准。

以下估计值由独立来源在开放科学协作组织的重复研究项目（2015年）之前确立。

统计功效 ($1-\beta \approx 0.35$)： Cohen（1962）记录了异常社会心理学中典型检验效能偏低的情况。在对同一份期刊进行后续历史比较时，Sedlmeier 和 Gigerenzer（1989年）报告称，中等效应的中位数功效持续偏低（1960 年为 0.46，1984 年为 0.37）；在将不显著结果解释为支持零假设的案例中，中位数功效低至 0.25。Button 等人（2013）报告称，神经科学中的中位数统计功效约为 21%，并总结指出，不同分析和子领域的功效大致介于 8% 至 31% 之间。我们采用 $1-\beta = 0.35$ 作为改革前社会心理学的中心估计值。

先验概率 ($\pi \approx 0.10$，校准参数)： 领域层面的先验无法直接观测，因此将其视为一个校准参数，用于概括进入正式检验的假设中真正为真的基准率。我们将 $\pi = 0.10$ 作为改革前社会心理学的中心值，这与 Ioannidis（2005）探讨的情景一致，并对 $\pi \in [0.05, 0.20]$ 进行敏感性分析。

显著性水平： $\alpha = 0.05$，即大部分文献中的常规阈值。

这些数值意味着 $\Lambda = 0.35/0.05 = 7$ 以及阳性预测值 $= 0.10 \times 7 /$（$0.10 \times 7 + 0.90$）$=$0.44。

将方框1中的参数代入Replication Bridge，重复研究功效为 $1-\beta_r \approx 0.75$（基于开放科学协作组织将重复研究设计为达到高检验效能的做法；参见开放科学协作组织，2015年）：[2]

[2] 赢家诅咒（Ioannidis，2008）会加剧这一效应：如果重复研究团队根据已发表的效应量 $\theta > \theta^{\hat{}}$ 来设定研究的检验效能，实际实现的重复研究功效就会低于预期水平。因此，在这些参数假设下，预测的重复率为 36–38%，应视为近似上界。

$$\mathbb{P}(\text{重复研究显著} \mid \text{原始研究显著}) = 0.44 \times 0.75 + 0.56 \times 0.05 = 0.330 + 0.028 = 0.358.$$

$$\text{With } 1-\beta_r = 0.80\text{: } \mathbb{P} = 0.44 \times 0.80 + 0.56 \times 0.05 = 0.380.$$

隐含的重复率为36–38%，与开放科学协作组织观察到的36%相一致（开放科学协作组织，2015年）。更精确的表述是回溯预测：将该框架应用于独立确立的参数范围，并将隐含比率与随后观察到的数值加以核对。桥接反演提供了补充检验：由 $R = 0.36$ 和 $\widehat{1-\beta_r} = 0.75$，我们得到阳性预测值 $= (0.36 - 0.05)/(0.75 - 0.05) = 0.44$，与由参数推导的估计相一致。这种吻合是框架内部的代数一致性，而非独立确证；其价值在于表明，在所观察到的重复率下，该桥接在内部是自洽的。

回溯预测并不能确定唯一的参数化。多种（$\pi, 1-\beta$）组合可以产生相似的重复率（表2），而桥接反演识别的是隐含的阳性预测值，而非单独的 π，且缺少关于原始统计功效和有效 α的假设。关键在于体制层面的稳健性：在广泛的、独立而言合理的改革前参数范围内，该领域仍低于目标可靠性的可行性边界。

表2在 $\alpha = 0.05$、 $\alpha_r = 0.05$、$1 - \beta_r = 0.75$条件下，不同先验概率和检验效能所对应的重复率。观测到的开放科学协作组织重复率*36%*以粗体显示。

	先验概率 π				
检验效能 $1-\beta$	0.05	0.10	0.15	0.20	0.30
0.25	20%	30%	38%	44%	53%
0.35	24%	**36%**	**44%**	**50%**	**58%**
0.50	29%	42%	50%	55%	62%

在全部十五个单元格中， $\Psi > 1$ 对于目标值 $\tau = 0.95$：不可行性结论是

对较大的参数不确定性具有稳健性。

其他一致性检验与该框架大体一致。Camerer 等（2016）发现实验经济学中的重复验证率约为61%；Camerer 等（2018）发现《自然》和《科学》中社会科学实验的重复验证率约为62%。这些较高的比率与较高的先验概率或较高的杠杆相一致。Errington 等（2021）报告称，临床前癌症生物学中的重复验证成功率因标准而异；桥接反演（假设 $\alpha_r = 0.05$ 且重复研究功效 ≈ 0.80）所还原的阳性预测值处于与不可行但多数为真的状态相一致的范围内。

5 导致崩塌的两条路径

可靠性需要足够的杠杆，但杠杆可能被削弱。科学实践中的两个反复出现的特征会推动杠杆下降，并且在崩塌极限下，可能使阳性预测值（PPV）回归先验：观察性数据中的持续混杂，以及发表压力下的分析灵活性。这两种机制都可能在标准学科实践中出现，并不要求存在恶意。坍缩结果是渐近的，并且具有机制特异性；这并不意味着所有观察性研究或所有设定灵活性在实践中都会坍缩至先验。

5.1 路径 I：观察性崩塌

定义 20（因果零假设与关联零假设）。令 $H_{\text{causal}} = 0$ 表示不存在因果效应，令 $H_{\text{assoc}} = 0$ 表示在调整已测量协变量后不存在关联。当存在大小为 $b > 0$ 的未测量混杂因素时，因果零假设并不意味着关联零假设：大小为 b 的虚假关联仍然存在。

在下述定理中，z_α 表示上 α尾标准正态临界值，即对于 $Z \sim N(0,\ 1)$，$\mathbb{P}(Z > z_\alpha) = \alpha$。

定理 21 (观察性研究坍缩**)**。考虑一项观察性研究，通过$Z_n = \theta_n / \mathrm{SE}(\theta_n)$检验因果效应，并对较大的正值进行单侧拒绝（双侧对应形式见*(d)*部分），且满足以下条件：

(i) 持续性未测量混杂：在被检验的方向上存在偏倚项 b_n

使得在 $H_{\text{causal}} = 0$下， Z_n 对某个 $\sigma > 0$渐近地为 $N(\sqrt{n}\, b_n/\sigma, 1)$，且$\sqrt{n}\, b_n \to +\infty$；

(ii) 不得以 n 的方式强化识别策略，从而使 $b_n \to 0$ 的速度快到足以阻止 $\sqrt{n}\, b_n \to \infty$；

(iii) 在 $H_{\text{causal}} = 1$ （$1 - \beta_n \to 1$）下的一致检验。
那么，当 $n \to \infty$时：

(a) α_{eff}（n） $\to 1$; *(b)* Λ_{eff}（n） $\to 1$; *(c)* 阳性预测值$_n \to \pi$; *(d)* 对于双侧拒绝（$|Z_n| > z_{\alpha/2}$），若 $\sqrt{n}\,|b_n| \to \infty$，则 α_{eff}（n） $\to 1$ 也成立。证明。见附录 11.3。

□

在所述假设下，观察性崩塌在性质上比固定-α 上限更严重。随机研究在 $\pi = 0.10$ 和 $\alpha = 0.05$下的上限为 69%；存在混杂的观察性研究的渐近值为 10%，即先验本身。当存在持续混杂时，增加样本量会使检验逐渐无法提供有关因果假设的信息。该定理适用于在固定识别策略下未测量混杂不会减弱的设计。如果不存在混杂，或者混杂通过改进的设计或测量而减弱，则未必会发生崩塌。若 $\sqrt{n}\, b_n \to -\infty$ 且检验方向为正，则 α_{eff}（n）为 $\to 0$ 而非 1，因此，混杂相对于拒绝域的方向性至关重要。

推论 22（连续估计下的确信性错误结论）。 在错误设定的 M 估计的标准正则条件下（例如 *White 1982*），如果拟合模型遗漏了导致偏差的混杂因素，那么 $\hat{}\theta_n \xrightarrow{p} \theta^\star$，其中 $\theta^\star$ 是伪真参数（错误设定的模型类中的 *Kullback–Leibler* 最小化解）；在常见的混杂情形下， $\theta^\star$ 与受混杂影响的关联估计目标一致。$\hat{}$此外，$\mathrm{SE}(\theta_n) \to 0$。在标准的贝叶斯模型设定错误正则条件下

（例如 *Kleijn* 与 *van der Vaart 2012*），后验同样会集中于 $\theta^\star$ 附近。二元坍缩使研究者重新面对不确定性（先验）；而连续估计坍缩则可能使其对结构性错误的估计目标产生高度确信。

这种不对称性有助于解释某些研究传统中一个原本令人费解的特征：大型观察性研究逐渐收敛于精确的估计值，而这些估计后来却被实验证据所反驳（Ioannidis, 2018）。文献不断累积以混杂偏倚为中心的紧窄置信区间，而不是诚实地反映不确定性。样本量降低了关于 $\theta^\star$的不确定性，却没有降低关于因果效应 θ的不确定性。

推论23（识别即逃逸）。 实现有效识别的设计（随机实验、工具变量、断点回归，以及在识别假设成立时的双重差分）会消除或界定 b，从而恢复名义杠杆。在定理 *21* 的假设下，如果不存在某种机制，迫使 $b_n \to 0$ 随着 $n \to \infty$ 以足够快的速度下降，那么有效识别（或等价的偏倚收缩机制）就是在该框架下实现渐近可靠因果推断的必要条件。

5.2 路径 II：规格搜索崩溃

定义24（序贯设定搜索）。 研究者拥有 $m \geq 1$ 个合法的分析设定。每个设定依次接受检验；如果结果显著，就报告该结果并停止搜索；如果结果不显著，则以 $q \in [0,1]$ 的概率发表零结果，或者继续检验下一个设定。参数q 控制选择性延续： $q = 0$ 意味着在穷尽搜索之前绝不发表零结果； $q = 1$ 意味着不存在选择性延续。

定义25（设定搜索下的有效错误率）。定义$\alpha_{\text{eff}}(q,\ m)$：$= \mathbb{P}$（一项显著结果被发表 $\mid H = 0$），以及$1 - \beta_{\text{eff}}(q,\ m)$：$= \mathbb{P}$（一项显著结果被发表 $\mid H = 1$）。

引理26（设定搜索下的有效错误率）。 在名义水平为（α、 $1 - \beta$）的 m 个独立检验下进行序贯搜索，定义 $s_0 =$ （$1 - \alpha$） （$1 - q$） 以及

$s_1 = \beta\,(1-q)$：

$$\alpha_{\text{eff}}(q,m) = \alpha \cdot \frac{1-s_0^m}{1-s_0}, \tag{11}$$

$$1-\beta_{\text{eff}}(q,m) = (1-\beta)\cdot\frac{1-s_1^m}{1-s_1}. \tag{12}$$

证明。 见附录 11.4。 □

这些公式在独立性基准下是精确的。独立性仅用于推导闭式解；坍缩端点（命题 28）只要求搜索能够在 $q=0$ 下使 $\alpha_{\text{eff}} \to 1$ 如 $m \to \infty$ 。不同设定之间的依赖性可能会减弱或放大这种膨胀，具体取决于相关结构。与分析灵活性相关的行为模型包括 Simmons 等人（2011）的研究，以及 Gelman & Loken（2014）对“分叉路径花园”的分析。

定理 27 (规格搜索下的判别损失**)**。 在参数为 (q, m) 的设定搜索下，有效杠杆为 $\Lambda^{\text{pb}}_{\text{eff}} = \Lambda \cdot D(q,m)$，其中上标***“pb”***表示发表偏倚设定，而判别损失因子为

$$D(q,m) := \frac{1-s_1^m}{1-s_0^m}\cdot\frac{1-s_0}{1-s_1} < 1 \tag{13}$$

对于所有 $q<1$、 $m \geq 2$ 和 $1-\beta > \alpha$。

证明。 见附录 11.5。 □

命题 28 （饱和与崩塌）。 假设 $\Lambda > 1$。

(a) 饱和 （固定 $q \in (0,1)$， $m \to \infty$）： 杠杆在 $\Lambda_{\text{sat}}(q) < \Lambda$处达到饱和；阳性预测值收敛到一个严格高于 π 、但低于无偏阳性预测值的数值。

(b) 无界搜索坍缩 *($q=0$, $m \to \infty$)*： $\Lambda^{\text{pb}}_{\text{eff}} \to 1$ 和 阳性预测值 $\to \pi$。

证明。 参见附录 11.6。 □

大多数真实领域可能处于饱和机制（$q>0$，有限的 m），而非字面意义上的崩塌极限。坍缩是揭示这一机制终点的极限情形；仅仅饱和就足以通过压低Λ_{eff} 至所需阈值以下，使一个领域处于不可行体制。

5.3 预注册推论

推论29 (预注册**)**。 在理想化合规条件下，预注册将主分析限定为单一的验证性设定（$m = 1$），从而恢复 $\alpha_{\text{eff}} = \alpha$，且不受 q影响。这消除了发表压力与设定多重性之间的关联，使阳性发现的阳性预测值不受文件抽屉强度影响。在实践中，部分合规或多项预先指定的分析可表示为 $m > 1$，并采用约束搜索。

5.4 双重坍缩

定理 30 (双重坍缩**)**。 假设观察性混杂和序贯设定搜索同时存在，且设定搜索参数 (m, q) 在n中固定。

（*a*）如果混杂使有效零假设拒绝率升至 *1*， $\alpha_{\text{eff}}(n, m, q) \to 1$，且在 $H = 1$ 下的有效检验效能保持一致， $1 - \beta_{\text{eff}}(n, m, q) \to 1$，那么$\Lambda_{\text{eff}}(n, m, q) \to 1$，当 $n \to \infty$ 时。

（*b*）仅无界搜索（$b = 0$、 $q = 0$、 $m \to \infty$）便会导致 $\Lambda_{\text{eff}} \to 1$。

证明。 参见附录 11.7。 □

在 (a) 部分中，混杂会驱动 $\alpha_{\text{eff}}(n,\ m,\ q) \to 1$；设定搜索可能已经将其推高至名义水平之上。贝叶斯更新因子（真阳性率与假阳性率之比）随后会变成两个趋于 1 的量之比，因此该商也趋于 1。这两种机制依次削弱贝叶斯更新：设定搜索会放大 α；混杂则会抹去剩余的杠杆。

示例 31 （双重坍缩：数值示例）。设 $\pi = 0.10$，检验效能为 $= 0.80$，名义 $\alpha = 0.05$，则 $\Lambda = 16$，阳性预测值为 ≈ 0.64。设定搜索会将 α_{eff} 推高至 0.20，使有效杠杆降至 4，阳性预测值降至约 0.31。加入观察性崩塌会使 Λ_{eff} 趋近于 1；当 $\Lambda_{\text{eff}} \to 1$时，阳性预测值为 $\to 0.10$。

这种双崩溃构型（观察性数据、灵活分析、强烈的发表压力）描述了历史上的候选基因文献（Border 等，2019）以及部分观察性营养流行病学研究（Ioannidis，2018）。在双崩溃构型中，单通道改革并不足够：更好的协变量调整无法阻止 α_{eff} 因设定搜索而膨胀，预注册也无法凭空创造有效识别。这两种做法都不意味着存在恶意。

6 逃离不可行状态

坍塌结果也指出了可用的逃逸路径。有三种机制可以修复杠杆，或推动一个领域走向可行体制。预注册（推论29）可确保名义 α，这是必要但通常不足的一步。其余两种机制——阈值收紧和重复验证——将在此展开讨论。其中，重复验证流水线的适用范围最广，因为它们直接关系到什么构成科学证据单元。

6.1 随机实验中的自适应逃逸

定理 32 (自适应逃逸**)**。 考虑针对简单备择的单侧高斯 z检验，其效应量为 $\theta_1 > 0$，已知标准误尺度为 $\sigma/\sqrt{n}$，且$\alpha_n = 1 - \Phi(c\sqrt{n})$，其中 $0 < c < \theta_1/\sigma$为某个常数。那么 $\alpha_n \to 0$ 呈指数级增长，$(1-\beta_n) \to 1$， $\Lambda_n \to \infty$，并且对于任意 $\pi \in (0,\ 1)$ 和 $\tau \in (0,\ 1)$，阳性预测值$_n \geq \tau$对所有充分大的 n均成立。所需样本量按 $n = O(\log \Lambda_{\text{req}})$增长。

证明。参见附录 11.8。 □

关键在于，自适应逃逸要求同时 提高 n 并收紧 α_n；在 $\alpha = 0.05$固定的情况下，单靠样本量无法将阳性预测值提高到固定-α 上限以上。最清晰的实证例子是全基因组关联研究，它对$\sim 10^6$ 变异实施Bonferroni校正（$\alpha = 5 \times 10^{-8}$；Pe’er 等，2008），将 Λ 从16提高到 1.6×10^7。从候选基因研究转向全基因组关联研究，是领域通过收紧阈值摆脱不可行体制的一个特别鲜明的例子；这种方法论上的断裂，其定量特征与库恩式范式转变相一致

（第8节）。

6.2 可重复性流水线：主要突破机制

即使是一项执行完美的预注册随机实验，单项研究在结构上也不足以让任何先验低于临界先验的领域达到目标可靠性。在这种情况下，自然的证据单位是研究流程。

定理 33（复现流水线）。 假设 $\Lambda > 1$。只有当所有 k项独立的预注册研究都达到显著性时，才接受某项主张；且每项研究均在相同的名义水平 α 下开展，检验效能为 $1-\beta$，则合并杠杆为

$$\Lambda^{(k)} = \left(\frac{1-\beta}{\alpha}\right)^k = \Lambda^k. \tag{14}$$

对于任意目标值 $\tau \in (0,\ 1)$ 和先验 $\pi \in (0,\ 1)$，存在 k^* 使得 阳性预测值$(\Lambda^{(k)}) \geq \tau$对所有 $k \geq k^*$均成立。

证明。 在给定 H的条件独立假设下，合并后的假阳性率为 α^k，合并后的真阳性率为 $(1-\beta)^k$，因此 $\Lambda^{(k)} = \Lambda^k$。由于 $\Lambda > 1$，我们有$\Lambda^k \to \infty$，因此根据定理7，阳性预测值$(\Lambda^{(k)}) \to 1$。因此，对于任意目标值 $\tau \in (0,\ 1)$，存在 k^* 使得阳性预测值$(\Lambda^{(k)}) \geq \tau$ 对所有 $k \geq k^*$均成立。

□

这一结果涉及证据标准（“当且仅当所有 k 研究都显著时才接受”）。如果只有成功的研究流程被选择性发表，那么研究流程层面的有效率 α 就会被高估，因此要保证这一结果，就必须使用有效率。在重复研究之间存在正相关的情况下，α^k 和 $(1-\beta)^k$ 不再是精确值；杠杆倍增变成了一个上界。

杠杆倍增具有几何级数效应：在 $\Lambda = 16$进行两次独立重复，得到 $\Lambda^{(2)} = 256$，足以在先验概率低至 $\pi \approx 7\%$ 时达到 $\tau = 0.95$。三次重复得到 $\Lambda^{(3)} = 4,\ 096$，足以在 $\pi \approx 0.5\%$ 时达到目标值。单项研究中，无论如何增加样本量都无法实现这一点，因为目标值尚未达到，固定-α 上限就已生效。在给定 H 的条件下，若满足条件独立，这种倍增效应是精确的（注15）；共享方法或总体会削弱这一效应。

推论34 (最小管线深度**)。** 若 $\Lambda > 1$，要达到阳性预测值 $\geq \tau$，全显著独立管线所需的最少研究数量（包括原始研究）为

$$k^* = \max\left\{1, \left\lceil \frac{\log \Lambda_{\text{req}}(\tau, \pi)}{\log \Lambda} \right\rceil \right\}.$$

等价地，超出原始研究的重复研究最少次数为 $k^* - 1$.

从制度层面看，其含义是直接的。一个领域的证据标准应当是一条预注册重复研究管线，而不是一项带有 p 值的单独研究。单个显著结果是证据的一部分，而不是一个证据单位。由此看来，将孤立发现视为可发表结论的期刊和资助方，正在将证据不足制度化。

7 领域动力学

本节将确定性界限纳入领域层面的演化模型，构建风格化的动态扩展。重点在于分析上的透明性和阈值行为，而非对特定领域的演变轨迹进行现实估算。

7.1 领域寿命定理

随着一个领域日趋成熟，其先验概率 $\pi(t)$ 往往会下降：真实关系被发现后，会从未解决问题的集合中移除；与此同时，竞争压力会促使推测性假设不断增加。在动力学 $\pi(t) = \pi_0 e^{-(\gamma+\delta)t}$ 下：

定理35 (领域寿命**)。** 假设 $\gamma + \delta > 0$, $\pi_0 > \pi_{\text{crit}}(\tau, \alpha, \beta)$，并且 α 和 β 保持不变。该领域在时间

$$T^*(\tau) = \frac{1}{\gamma + \delta} \ln\left(\frac{\pi_0}{\pi_{\text{crit}}(\tau, \alpha, \beta)} \right). \tag{15}$$

对于 $t > T^*$，目标阳性预测值在结构上无法达到。

证明。 令 $\pi_0 e^{-(\gamma+\delta)t} = \pi_{\text{crit}}$ 并求解。 □

示例36 （阈值收紧对领域寿命的效应）。 一个领域具有$\pi_0 = 0.70$、 $\gamma + \delta = 0.05$、检验效能为 $= 0.80$、目标值为 $\tau = 0.95$。在 $\alpha = 0.05$时： $\pi_{\text{crit}} = 0.543$、$T^* \approx 5$年。在 $\alpha = 0.005$时： $\pi_{\text{crit}} = 0.106$、 $T^* \approx 38$年。将 α降低十倍，可使有效寿命延长七倍。

7.2 G 代际动态与退化方案准则 n

当后续研究建立在已发表研究发现之上时，某一代中的假阳性会削弱下一代可用的先验概率。

定理37 (代际动态**)**。 令 阳性预测值$_k$ 表示第 k 代的阳性预测值，其中后续假设以父级发现为真的概率为$\pi_c \in (0,1)$为条件。第 $k+1$ 代的有效先验为 $\pi_{k+1} =$ 阳性预测值$_k \cdot \pi_c$，阳性预测值的演化如下

$$\mathrm{PPV}_{k+1} = \frac{\mathrm{PPV}_k\, \pi_c \Lambda}{\mathrm{PPV}_k\, \pi_c \Lambda + (1 - \mathrm{PPV}_k\, \pi_c)}. \tag{16}$$

(a) 坍缩 *(*$\pi_c\Lambda \le 1$*)*：阳性预测值$_k \to 0$ 单调变化。在边界情形 $\Lambda = 1$ 下，阳性预测值$_{k+1} =$ π_c阳性预测值$_k$，因此 阳性预测值$_k = \pi_c^k$ 阳性预测值$_0 \to 0$ ，对于 $\pi_c < 1$。

(b) 恢复 *(*$\pi_c\Lambda > 1$*)*：如果 阳性预测值$_0 > 0$，那么 阳性预测值$_k \to (\pi_c\Lambda - 1)/[\pi_c(\Lambda - 1)]$。

证明。 参见附录11.9。 □

从代数上看， $\pi_c\Lambda \le 1$ 与应用于后续研究的多数为假阈值相同，为拉卡托斯（1978年）关于退化研究纲领的概念提供了定量桥梁。将 $\pi_c\Lambda$ 称为该纲领的**进步比率**。

推论38 （退化研究纲领判据）。 如果一个建立在已发表研究发现之上的研究纲领可证明地退化 （产生可靠性逐代下降并趋向噪声的连续世代），当且仅当其进步比率满足$\pi_c\Lambda \le 1$。

退化纲领已使自身后继研究处于多数为假状态。当进步比率降至1或更低时，恢复需要提高 π_c （将后续研究限制在更高质量的候选对象上），或提高 Λ （采用更严格的阈值或可重复性流水线）。如果不进行这些改变，每一代都会继承更弱的先验，而文献则会逐渐趋向噪声。

例39 （坍缩与恢复）。 坍缩。 使用 $\Lambda = 7$、$\pi_c = 0.10$开展候选基因研究：$\pi_c\Lambda = 0.7 < 1$。

代数 k	π_k	PPV_k	每个真阳性对应的假阳性数
0	0.020	0.125	7.0
1	0.013	0.081	11.3
2	0.008	0.054	17.4
3	0.005	0.037	26.2

恢复。 使用 $\Lambda = 1.6 \times 10^7$、 $\pi_c = 0.50$ 的全基因组关联研究： $\pi_c\Lambda \gg 1$。阳性预测值迅速恢复。

7.3 自我强化型退化

一旦形成，这种反馈就会不断叠加。假阳性会引发推测性后续研究，降低下一代的 π ；下降的阳性预测值又会产生更多假阳性，进一步降低 π 。这条链条中的每个环节，都是将确定性界限应用于劣化先验的实例。该机制并不要求不当行为增加；它只要求后续研究将已发表研究发现视为有关下一步应去何处寻找的有用信息。

处于这种状态的领域无法仅通过渐进式改进实现自我纠正。除非提高 π_c 或实质性增加有效杠杆 Λ，否则更大的样本和更严谨的统计实践都无法遏制衰退。在此处的程式化动力学中（固定的 Λ 和固定的 π_c），这些干预不会改变结构参数。因此，要摆脱这种状态，就需要外生地改变其中一个参数。

候选基因文献并非通过逐步完善而恢复，而是需要不连续地采用全基因组关联研究（Border 等，2019），使 Λ 提高了六个数量级。粒子物理学的 5σ 阈值（Cowan 等，2011；ATLAS 合作组，2012；CMS合作组，2012）以及注册报告的采用（Chambers，2013），代表了类似的外生干预。退化性研究纲领判据不仅揭示了哪些领域陷入困境，也说明了为何恢复需要改变运行参数的外生变化，而不是在现有范式内进行渐进式完善。

说明 40（动力学模型的适用范围）。 自我强化型退化模型经过了简化：它假设 Λ 固定，并假设存在一条确定性传递规则 $\pi_{k+1} = \text{阳性预测值}_k \cdot \pi_c$。

真实领域中存在随机变化、部分更新和异质性子研究计划。该模型的价值在于识别定性阈值（$\pi_c \Lambda \leq 1$），低于该阈值时，漂移方向在结构上是确定的；而不在于对特定领域轨迹作出定量预测。

8 可靠性景观

8.1 可靠性景观与库恩式转变

可靠性景观由两个结构参数定义：一条轴上的实验杠杆 Λ，以及另一条轴上的先验概率 π。目标值 τ 的可行性边界是曲线 $\pi\Lambda = \tau(1-\pi)/(1-\tau)$，等价地表示为 $\Psi = 1$。位于这条曲线上方的领域处于可行体制；位于曲线下方的领域则处于不可行体制。

在这一景观中，常规科学处于固定位置。渐进式改进（样本略有增加、测量略有改善）会带来连续而细微的移动。然而，方法论的根本转变会使 Λ 发生不连续的向右跃迁，一步之内便可能跨越可行性边界。全基因组关联研究的转变使 Λ跃升了六个数量级；粒子物理学采用 5σ 标准（Cowan等人，2011年；ATLAS 合作组，2012；CMS合作组，2012），则以类似方式说明了阈值收紧如何成为一种制度化证据标准。两者都不可能通过渐进式改进实现；都必须取代证据标准本身。

跨越可行性边界的运动捕捉到了库恩（1962）所称范式转变的定量特征。解释的方向至关重要：跨越边界并不定义 范式转变，但方法论革命往往会产生它们。库恩定性描述为革命的现象，在这一框架中具有可测量的特征：杠杆的不连续增长跨越了 $\Psi = 1$边界。

8.2 领域校准

随着一个领域逐渐成熟，π 会下降，该领域也会沿纵轴向下移动。若没有相应收紧阈值（向右移动），就会越过可行性边界。表3展示了代表性领域的可靠性景观。

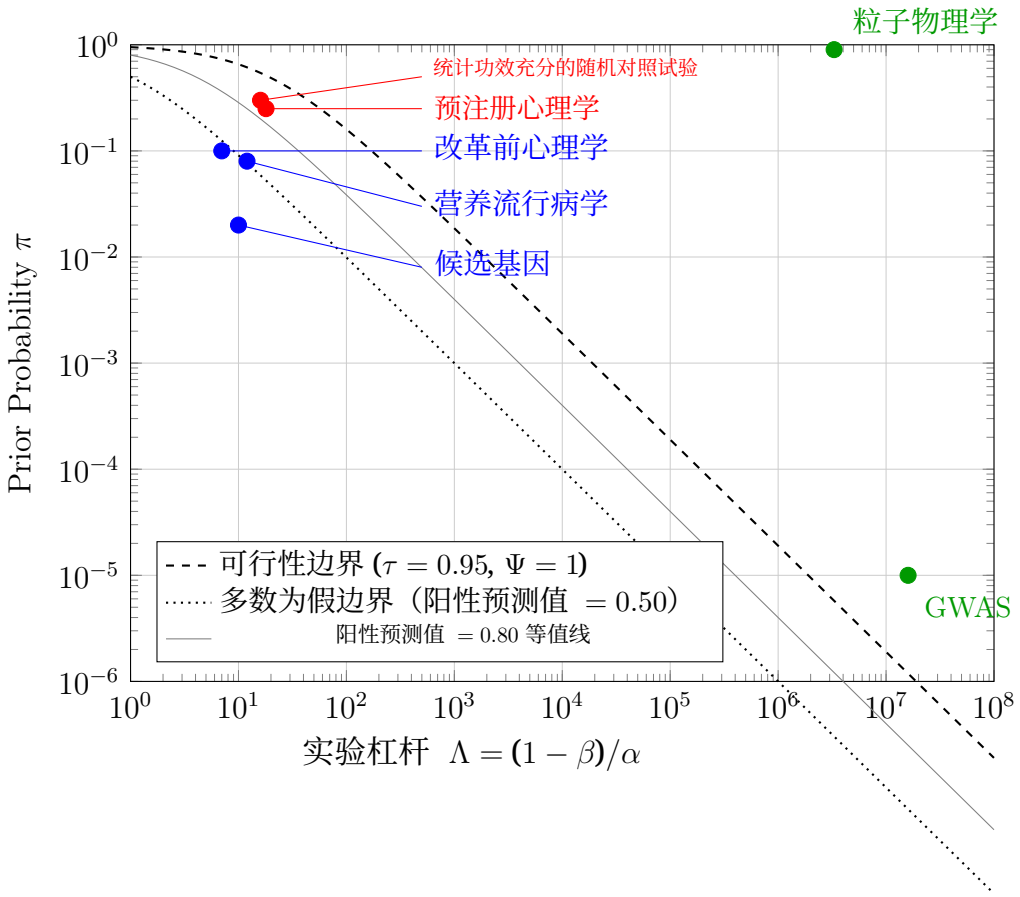


图2

科学领域的可靠性景观。各点表示示意性校准，而非精确测量；其目的是展示不同状态。虚线是 $\tau = 0.95$ ($\Psi = 1$)的可行性边界：位于其上方且右侧的领域是可行的。点线是多数为假边界。蓝色：多数为假的示例。红色：虽有改善但仍不可行。绿色：可行。

表 3 的目的是在合理的参数化方案下展示状态，而非对各领域进行精确估算。

即使经过预注册的心理学研究，在 $\Psi \approx 3.2$ 仍处于不可行体制，这表明仅靠预注册并不足够，还需要更严格的阈值或显式的重复验证要求。

9 可靠研究的结构性要求

对于在基于二元显著性的发表架构下以声明的可靠性目标值 τ 为目标的领域，要处于可行范围内，必须满足以下四个条件

表3目标值为 $\tau = 0.95$ 时代表性领域的状态映射情景示例。

领域	α	检验效能	π	Λ	Ψ	PPV	上限	状态
候选基因	0.05	0.50	0.02	10	93	17%	29%	多数为假
改革前心理学	0.05	0.35	0.10	7	24	44%	69%	不可行
营养流行病学	0.05	0.60	0.08	12	18	51%	63%	不可行
统计功效充分的随机对照试验	0.05	0.80	0.30	16	2.8	87%	90%	不可行
预注册心理学	0.05	0.90	0.25	18	3.2	86%	87%	不可行
GWAS	5×10^{-8}	0.80	10^{-5}	1.6×10^{7}	0.12	99%	∼100%	可行
粒子物理学	3×10^{-7}	0.9999	0.90	3.3×10^{6}	6×10^{-7}	∼100%	∼100%	可行

注。 所有 π 值均为示例性校准值，并非精确估计。对于改革前心理学，$\pi = 0.10$ 是一个校准参数，与 Ioannidis（2005）中的情景及第4节中的回溯预测一致。候选基因的校准值设定为与文献中有充分记录的失败情况相符（Border 等，2019）；这一行是最稳健的结果之一，因为文献中的失败情况已得到充分

已有记载。营养流行病学数值借鉴Ioannidis（2018），仅为示意；不可行性分类在 $\pi \in [0.03, 0.20]$ 范围内均稳健。随机对照试验一行最为敏感：仅当 $\pi \gtrsim 0.54$ 在所述参数下时才进入可行。预注册心理学（$\pi \approx 0.25$）反映这样一种推测，即预注册伴随着更多理论驱动、基准率更高的研究；这是最不确定的估计。全基因组关联研究参数反映针对 $\sim 10^6$ 个变异的Bonferroni校正（Pe’er 等，2008）。粒子物理学使用说明性的 5σ水平阈值（单侧 $\alpha \approx 3 \times 10^{-7}$）（Cowan

等，2011年）。对 π 的敏感性在表2中加以探讨。

状态。在此框架中，这些是由数学推导出的设计要求，而非可自行决定的指导方针。

1. 阈值校准。领域的显著性阈值必须满足

$$\alpha \le \alpha_{\max}(\pi, \beta, \tau) = \frac{(1-\beta)(1-\tau)\pi}{\tau(1-\pi)}. \tag{17}$$

对于 $\pi = 0.10$、 $\tau = 0.95$，检验效能为 0.80： $\alpha_{\max} = 0.0047$。Benjamin 等人（2018）的提议（$p < 0.005$）符合这一要求。

2. 名义水平的执行。 预注册和注册报告强制执行 $\alpha_{\text{eff}} \approx \alpha$ （推论29）。名义 α 只有在有效 α_{eff} 为

受控。

3. 将重复验证作为证据单位。 在常规 $\alpha = 0.05$ 和目标值$\tau = 0.95$下，固定-α 上限低于目标值，除非满足 $\pi \gtrsim 0.49$。对于许多远低于这一阈值运行的领域，标准因此必须是一条复现流水线： k 项独立研究产生杠杆效应 Λ^k （定理 33）。科学发现是这条流水线的产出。

4. 对因果主张进行有效识别。 无论阈值或样本量如何，都无法克服观察性崩塌（定理 21）。观察性研究可以为关联提供有用证据；需要有效识别的，是将显著关联解释为因果效应。

并非所有改革都同样有效。预注册控制 α_{eff} ，但不会改变 π 或名义上的 Λ。阈值收紧会直接增加 Λ。重复验证会使 Λ 呈几何级数增长。在固定 α 下增加样本量，仍受上限约束。若“改进实践”的倡议不改变运行参数（π、 α、 $1-\beta$），就无法改变 Ψ ，因此本身也无法推动某个领域跨越可行性边界。

9.1 一个实例：临床前阿尔茨海默病研究

为具体说明这些要求，可将临床前阿尔茨海默病文献作为一个程式化案例来考察。假设某一临床前靶点假设正确的先验概率为 $\pi \approx 0.05$；鉴于中枢神经系统药物开发的高淘汰率，这一取值可以说偏于乐观（见 Cummings 等，2014）。典型的临床前研究往往在 $\alpha = 0.05$ 的水平上运行，而在相邻的临床前文献与神经科学文献中，偏低至中等的检验效能十分常见（Button 等，2013）；我们在此以检验效能 ≈ 0.50 作为示例值。

在这些参数下： $\Lambda = 10$，阳性预测值 $= 0.34$， $\Psi = 36$，且固定-α 上限为 51%。该领域处于多数为假状态。其显著的临床前发现中，预计有一半以上为假阳性；这并非因为研究者能力不足，而是由运行参数所致。

如何补救？将 $\alpha = 0.005$ 收紧至 Λ ，可使阳性预测值升至 100，且

升至 84%，但 Ψ 仍高于 1。采用 $k = 2$ 次独立重复验证的研究流程，在$\alpha = 0.005$ 时可得到 $\Lambda^{(2)} = 10,000$，阳性预测值为 $> 99\%$。作者可以在设计阶段计算 Ψ；期刊可以要求将其与标准功效分析一并披露；而缺乏重复验证计划的 $\Psi > 1$，则可视为需要修改的设计局限。

更广泛地说，最低限度的证据状态报告应包括：目标可靠性 τ；假定的先验范围； α、检验效能和 Ψ；计划的重复验证深度 k；以及因果主张的识别状态。

10 讨论

10.1 与既有工作的关系

Ioannidis（2005）指出，在现实假设下，许多已发表研究发现都是假阳性。本文沿着这一论证思路，从三个方面加以拓展：从诊断转向结构性不可行，从静态快照转向领域动力学，以及从分析性批评转向连接该框架与观测到的重复率的经验桥梁。表4总结了这些区别。

表4与既有方法的比较。

	Ioannidis（2005）	**选择模型**	**诊断工具**	**本文**
Core 问题	假阳性的频率有多高?	真实效应量是多少?	真实信号?	可靠性能否实现?
阳性预测值上限	隐式	未设为目标值	未设为目标值	显式
坍缩	None	出版物筛选器	None	两种坍缩途径已证实
动力学	None	None	None	退化判据
哲学	None	None	None	波普尔、库恩、拉卡托斯

关于发表偏倚的既有研究涵盖选择模型（Hedges，1984；Andrews & Kasy，2019）和行为模型（Simmons 等，2011；Gelman & Loken，2014）。选择模型用于估计经偏倚校正的效应量；确定性界限则探讨在给定选择环境下，阳性预测值在结构上能够达到什么水平。两者并不冲突：可以使用 Andrews-Kasy 校正估计值重新校准 π 和检验效能，然后将确定性界限应用于校正后的参数。

有必要对三种诊断工具进行具体比较。*P* 曲线 （Simonsohn 等，2014年）检验一组显著 p值是否具有证据价值；*z* 曲线 （Bartoš & Schimmack，2022年）估计预期重复验证率和发现率。二者都评估现有文献中的证据内容。确定性界限提出了一个互补性问题：鉴于其设计参数，该文献能否包含可靠信号？某篇文献可能通过 p曲线检验（效应是真实的），同时仍处于不可行体制中（阳性预测值过低，无法达到可靠性目标）。

10.2 与科学哲学的联系

下文阐述的哲学联系是解释性的桥梁，并非主张形式体系与历史上的哲学论述之间完全等价。其目的不是将波普尔、库恩或拉卡托斯简化为单一指标，而是说明，该框架为这些传统以定性方式描述的可靠性约束提供了一种定量结构。

10.2.1 波普尔、*Mayo* 与严厉性标准

波普尔（1959年）认为，经确证的假设（即经受住严肃证伪尝试的假设）值得暂时接受。确定性界限通过明确何时有理由在认识论上接受此类假设，进一步细化了这一推论：当检验程序相对于先验具有足够的杠杆作用时，便是如此，而这可用不可行性比率 Ψ 来概括。

与 Mayo（1996）的严厉性标准尤为密切。Mayo 的误差统计框架认为，只有当某个假设通过严厉检验时，才能说它通过了检验；而这要求：如果错误确实存在，该程序有很高的概率将其检测出来——低 α （没有信号时，该程序很少发出信号），高 $1-\beta$ （该程序能够可靠地检测真实

效应）。杠杆 $\Lambda = (1-\beta)/\alpha$ 是对这种类似严厉性之区分度的定量表达；它将这两个组成部分合并为一个比率，用以衡量显著结果在真实情形下比在零假设下可能出现的概率高出多少。

“类严格性”这一限定是有意为之：Mayo的完整严格性概念还涉及所探测错误的特异性，而这一维度无法仅由标量 Λ 单独捕捉。波普尔要求确证性检验必须严格，但缺乏一套形式装置来规定有多严格；Mayo提供了概念框架；确定性界限则给出一个阈值。一项检验满足目标可靠性 τ 的这一类严格性标准，当且仅当 $\Psi < 1$。低于该阈值时，一个达到显著性的假设在此意义上并未经过严格检验：给定先验，该程序无法实现所声称的可靠性。

在多数为假状态中，即便一个假设在 $\alpha = 0.05$处以合理的检验效能达到显著性，它更可能是假阳性，而非真实效应。证伪在逻辑上仍然有效，但其认识收益受到杠杆的约束。

10.2.2 库恩与范式转变

在这一图景（图2）中，常规科学在固定位置上运作；范式转变则是跨越 $\Psi = 1$ 边界的不连续跳跃。跨越边界并不是范式转变的定义；它是一种可测量的定量标志，往往由方法论革命产生。全基因组关联研究的转变以及粒子物理学中的5σ 判据是典型例证：这些领域的跨越并非通过改进单项研究，而是通过改变什么算作证据。关于更广泛的确证理论背景（包括贝叶斯争论），参见Earman（1992年）；关于更宽泛的科学哲学综述，参见Gillies（1993年）。

10.2.3 拉卡托斯与退化研究纲领

退化性研究纲领判据（推论38）为拉卡托斯（1978年）对进步与退化研究纲领的区分设定了一个定量阈值：$\pi_c \Lambda \leq 1$。该判据捕捉的是拉卡托斯式退化的可靠性维度，即研究产出趋于噪声这一经验特征，而非其完整概念；完整概念还涉及理论停滞以及对辅助

假设的事后调整。当 $\pi_c \Lambda \leq 1$ 时，恢复需要改变外生运行参数，而不是在范式内部付出渐进式努力。候选基因到全基因组关联研究的转变，既是库恩意义上的范式转变，也是拉卡托斯意义上摆脱退化的过程；确定性界限为这两种情形都提供了阈值。

10.3 信息性悖论

分析得出的一个推论与通常的直觉相悖：在多数为假状态下，非显著结果比显著结果更具信息性。

命题41 (零结果信息性**)。** 如果检验具有区分度（$1-\beta > \alpha$），且阳性预测值 $< 1/2$，那么阴性预测值 $> 1/2 >$ 阳性预测值：与显著结果相比，非显著结果是更可靠的真实状态指示指标。

证明。 参见附录 11.10。 □

对于改革前心理学（$\alpha = 0.05,\ \beta = 0.65,\ \pi = 0.10$），模型给出的阴性预测值为 $= 0.93$，阳性预测值为 $= 0.44$。在这种参数化下，零结果有 93% 的可能性是正确的，而显著结果正确的可能性只有 44%。因此，筛选阳性发现的期刊，优先选择了可靠性较低的结果类别。在多数为假状态下，文件抽屉隐藏的不只是“零结果”，还隐藏了可靠性。

由此还可得出两个推论。第一，出版物筛选器会产生误导性文献：在已发表的显著发现中，假阳性多于真阳性。第二，在固定 α 的情况下，错误信息率存在一个无法通过增加样本量来降低的下限：

$$1 - \mathrm{PPV} \ \geq \ 1 - \mathrm{PPV}^{\mathrm{ceil}} = \frac{\alpha(1-\pi)}{\pi + \alpha(1-\pi)}. \tag{18}$$

在低于多数为假阈值的领域，预注册并发表零结果，将使文献转向信息量更高的部分。已发表文献的可靠性之所以会提高，并非因为开展了更多研究，而是因为在这种状态下，出版物筛选器将不再压制更可靠的结果类别。

10.4 荟萃分析

如果各项纳入研究的有效杠杆都接近 1，且存在相同的识别失败，那么荟萃分析也可能继承这种崩塌：汇总会围绕一个有偏估计目标不断提高精度，而非恢复有效识别。在崩塌极限下，每个低杠杆输入所提供的证据区分度都很小，因此单靠汇总无法恢复识别。当纳入研究保持足够的杠杆和有效识别时，荟萃分析仍是一种强有力的工具；这种崩塌继承现象仅适用于存在共同失效模式的低杠杆输入。因此，将荟萃分析置于顶端的证据等级体系，应以纳入研究的杠杆和识别质量为条件，而不应仅仅依据其数量。

10.5 局限性与开放问题

先验概率无法直接观测，因此校准依赖于元科学。不过，该框架并不要求精确的 π：它规定了在给定检验配置下，π必须达到何种程度，才能实现所声称的可靠性。独立性假设是一项建模基准（注15）；研究之间的依赖性可能会以任一方向改变有限样本杠杆。

该分析针对二元显著性主张，这是对任何研究中可获得的完整证据进行的有意简化。连续估计和贝叶斯方法能够提取更多信息；当前结果刻画了二元发表架构所施加的上限。该框架并不声称所有领域都具有单一的 π，所有研究彼此独立，或连续证据都以这种方式受到限制。在给定运行参数的条件下，它为二元显著性架构提供了可行性约束。文中的先验 π 始终指被选中进行检验的假设为真的概率，这是该领域假设生成过程的属性，而不是所有可设想假设中为真的比例。

架构层面和行为层面的解释在分析上彼此不同，但在经验层面相互关联：激励结构会降低 π，而注册报告等行为改革（Chambers，2013）会强制执行 $\alpha_{\text{eff}} \approx \alpha$，并可能通过改变人们选择研究哪些假设，间接提高 π

所研究的假设。架构框架刻画了这些约束；行为改革则是促使研究在这些约束范围内调整的一种机制。

由此自然引出三个未决问题：通过跨领域桥接反演对 π 进行经验估计、非对称误差代价结构，以及发表压力下序贯设计的阳性预测值性质。

11 结论

在基于二元显著性的发表架构中，确定性界限意味着可重复性危机的一个重要组成部分是结构性的。开放科学协作组织项目之前所记录的改革前心理学参数范围，意味着重复率约为36%，与观测值一致。据此，提高可靠性需要结构性改革：将阈值校准至领域先验，通过预注册或注册报告强制执行名义显著性水平，将可重复性流水线制度化作为证据单位，并为观察性研究中的因果主张确保有效识别。

该框架与科学哲学的联系是实质性的，而非装饰性的。它给出了波普尔式证伪在认识论上具有信息量的定量条件，为库恩式方法论革命在可靠性景观中提供了可测量的标志，并为拉卡托斯式退化研究纲领提供了可靠性阈值。

实际方案很直接：在设计阶段计算 Ψ ，将其与标准功效分析一并披露，在 $\Psi > 1$ 的情形下要求复现流水线标准，并在识别尚未确立时区分关联主张与因果主张。在此框架内，这些并非可选的改进，而是基于二元显著性的证据之数学所蕴含的设计要求。

在二元显著性架构下，无论研究者多么勤勉，忽视这些要求的领域都无法达到其所声称的可靠性。这并非道德谴责，而是设计诊断。

数据、代码和材料可用性

本文主要是理论性的，不报告对原创数据集的分析。结果、图表和表格均依据文中所述公式和参数设置得出。一个实现了Ψ诊断、复现流程计算器和可靠性景观可视化工具的交互式网页工具可访问：https://mpollanen.github.io/certainty-bound-tool/（存档于 https://osf.io/c5wun）。该交互式工具的源代码可通过存档项目获取，并采用知识共享署名 4.0 国际许可协议。该工具完全在客户端运行；不会传输任何数据。

利益冲突与资助

作者声明不存在利益冲突。作者感谢加拿大自然科学与工程研究委员会（NSERC）的资助，资助编号为 RGPIN-2019-04085。

作者贡献

Marco Pollanen：概念化、正式分析、调查、方法论、撰写——初稿、撰写——审阅与编辑。

附录

*

附录：次要结果的证明

11.1 命题 13 的证明（先验异质性惩罚）

设 PPV(π) $= \pi\Lambda/[\pi(\Lambda-1)+1]$。求二阶导数：PPV$''$($\pi$) $= -2\Lambda(\Lambda-1)/[\pi(\Lambda-1)+1]^3 < 0$，其中 $\Lambda > 1$，由此可知其严格凹。根据詹森不等式，$\mathbb{E}$[PPV(Π)] <PPV($\bar{\pi}$)（当 Π 非退化且$\Lambda > 1$时为严格不等式）。在 $\bar{\pi}$ 附近进行二阶泰勒展开，得到近似式 PPV($\bar{\pi}$) $- \mathbb{E}$[PPV(Π)] $\approx \Lambda(\Lambda-1)\sigma_\pi^2/[\bar{\pi}(\Lambda-1)+1]^3$。

□

11.2 定理 17 的证明（发现成本）

在独立同分布假设下，每项研究以概率$\pi(1-\beta)$得到一个真阳性，以概率 $(1-\pi)\alpha$ 得到一个假阳性。在 N 项独立研究中，预期数量分别为 $N\pi(1-\beta)$ 和 $N(1-\pi)\alpha$，因此它们的比值为 $(1-\pi)\alpha/[\pi(1-\beta)]$。直接代入阳性预测值公式可知，这等于 (1 − 阳性预测值)/阳性预测值。

□

11.3 定理 21（观察性崩塌）的证明

(a) 在 $H_{\text{causal}} = 0$ 下：根据假设，Z_n 渐近为 $N(\sqrt{n}\, b_n/\sigma,\ 1)$。由于 $\sqrt{n}\, b_n/\sigma \to \infty$，$\alpha_{\text{eff}}(n) = \mathbb{P}(Z_n > z_\alpha \mid H_{\text{causal}} = 0) \to 1$。

(b) 在 $H_{\text{causal}} = 1$ 下：根据一致性，$1 - \beta_n \to 1$，因此 $\Lambda_{\text{eff}} = (1-\beta_n)/\alpha_{\text{eff}}(n) \to 1$.

(c) 根据确定性界限：阳性预测值 $= \pi\Lambda_{\text{eff}}/[\pi\Lambda_{\text{eff}} + (1-\pi)] \to \pi$。

(d) 对于双侧拒绝（$|Z_n| > z_{\alpha/2}$）：若 $\sqrt{n}\,|b_n| \not\to \infty$，则在 $H_{\text{causal}} = 0$ 下依概率 $|Z_n| \to \infty$，因此 $\alpha_{\text{eff}}(n) \to 1$。

□

11.4 引理26的证明（有效错误率）

在 H_0 下，第 k 次尝试达到显著性的概率为 $\alpha \cdot s_0^{k-1}$，其中 $s_0 = (1-\alpha)(1-q)$。对 $k = 1$、.....、m 求和：$\alpha_{\text{eff}} = \alpha(1-s_0^m)/(1-s_0)$。以 $s_1 = \beta(1-q)$ 进行类似推导，可得有效功效公式。

□

11.5 定理 27 的证明（区分损失）

将有效率代入确定性界限：

$$\Lambda_{\text{eff}}^{\text{pb}} = \frac{(1-\beta)\frac{1-s_1^m}{1-s_1}}{\alpha\frac{1-s_0^m}{1-s_0}} = \Lambda \cdot \underbrace{\frac{1-s_1^m}{1-s_0^m} \cdot \frac{1-s_0}{1-s_1}}_{D(q,m)}.$$

定义 $h(s) := \frac{1-s^m}{1-s} = \sum_{j=0}^{m-1} s^j$。由于 h 在 $[0,1)$ 上严格递增，且 $s_1 < s_0$（因为 $1-\beta > \alpha$ 意味着 $\beta < 1-\alpha$，所以 $\beta(1-q) < (1-\alpha)(1-q)$），我们有 $h(s_1) < h(s_0)$，从而得到 $D = h(s_1)/h(s_0) < 1$。

□

11.6 命题 28 的证明（饱和与崩塌）

(a) With $q \in (0,1)$: $s_0, s_1 \in (0,1)$, so $s_0^m, s_1^m \to 0$. In the limit,

$D(q, \infty) = (1-s_0)/(1-s_1)$ 和 $\Lambda_{\text{sat}} = \Lambda \cdot D(q, \infty)$。这满足 $1 < \Lambda_{\text{sat}} < \Lambda$，因为$q[(1-\beta) - \alpha] > 0$。

（b）当 $q = 0$ 时：$s_0 = 1-\alpha$，$s_1 = \beta$。由于 $m \to \infty$：$\alpha_{\text{eff}} \to 1$ 和 $1-\beta_{\text{eff}} \to 1$。因此 $\Lambda^{\text{pb}}{}_{\text{eff}} \to 1$，且阳性预测值为 $\to \pi$。

□

11.7 定理30的证明（双重坍缩）

(a) 在所述假设下，$\alpha_{\text{eff}}(n,m,q) \to 1$ 且 $1-\beta_{\text{eff}}(n,m,q) \to 1$，因此 $\Lambda_{\text{eff}}(n,m,q) = (1-\beta_{\text{eff}}(n,m,q))/\alpha_{\text{eff}}(n,m,q) \to 1$。

(b) 仅在设定搜索下：由命题 28(b)，两个速率均为 $\to 1$，因此$\Lambda_{\text{eff}} \to 1$。

□

11.8 定理 32 的证明（自适应逃逸）

取 $\alpha_n = 1 - \Phi(c\sqrt{n})$：由于 $c\sqrt{n} \to \infty$，$\alpha_n \to 0$。在 H_1 下，$Z_n \sim N(\theta_1\sqrt{n}/\sigma, 1)$，因此 $1-\beta_n = \Phi((\theta_1/\sigma - c)\sqrt{n}) \to 1$，因为 $c < \theta_1/\sigma$。从而 $\Lambda_n \to \infty$。

$\Lambda_n \geq \Lambda_{\text{req}}$所需的样本量约为 $n \approx (2/c^2)\ln \Lambda_{\text{req}}$，因为 α_n呈指数衰减。

□

11.9 定理37的证明（代际动态）

将 $\pi_{k+1} =$ 阳性预测值$_k$ π_c 代入确定性界限，得到阳性预测值$_{k+1} = f($阳性预测值$_k)$，其中 $f(x) = x\pi_c\Lambda/(x\pi_c\Lambda + 1 - x\pi_c)$。

不动点。设定 $f(x) = x$：要么 $x = 0$，要么 $x^* = (\pi_c\Lambda - 1)/[\pi_c(\Lambda - 1)]$，，其中正不动点存在当且仅当 $\pi_c\Lambda > 1$。

坍缩（$\pi_c\Lambda \le 1$）。分子因子 $g(x) = \pi_c\Lambda - 1 - x\pi_c(\Lambda - 1)$ 满足$g(x) \le 0$，对所有 $x \in (0, 1]$, 无论 $\Lambda \ge 1$ 还是 $\Lambda < 1$ 均成立。因此 $f(x) \le x$，该序列单调不增，且下界为 0，唯一的不动点为 $x = 0$。在边界情形 $\Lambda = 1$ 中， $f(x) = \pi_c x$，因此对于 $\pi_c < 1$，$\mathrm{PPV}_k = \pi_c^k \, \mathrm{PPV}_0 \to 0$ 。

恢复 ($\pi_c\Lambda > 1$)。对于 $x \in (0, x^*)$: $f(x) > x$（趋向 x^*）。对于$x > x^*$: $f(x) < x$ （趋向 x^*）。根据单调收敛，阳性预测值$_k \to x^*$。

□

11.10 命题41（零结果信息性）的证明

阴性预测值 $= (1 - \pi)(1 - \alpha)/[(1 - \pi)(1 - \alpha) + \pi\beta]$。阳性预测值 $< 1/2$ 当且仅当 $\pi < \alpha/[(1 - \beta) + \alpha]$，阴性预测值 $> 1/2$ 当且仅当 $\pi < (1 - \alpha)/[(1 - \alpha) + \beta]$。第一个阈值低于第二个阈值，当且仅当 $\alpha[(1 - \alpha) + \beta] < (1 - \alpha)[(1 - \beta) + \alpha]$，这可简化为 $1 - \beta > \alpha$（判别条件）。因此，每当阳性预测值 $< 1/2$，阴性预测值也满足 $> 1/2$。

□

12 参考文献

Andrews, I., & Kasy, M.（2019）。发表偏倚的识别与校正。《美国经济评论》，*109*(8), 2766–2794。https://doi.org/10.1257/aer.20180310

ATLAS 合作组。（2012）。利用大型强子对撞机上的 ATLAS 探测器，在搜寻标准模型希格斯玻色子的过程中发现了一种新粒子。《物理快报*B*辑》，*716*(1)，1–29。https://doi.org/10.1016/j.physletb.2012.08.020

Bartoš, F., & Schimmack, U.（2022）。Z 曲线 2.0：估算重复率和发现率。*Meta-Psychology*，*6*，文章编号 MP.2021.2720。https://doi.org/10.15626/MP.2021.2720

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Johnson, V. E.（2018）。重新界定统计显著性。《自然·人类行为》，*2*(1)，6–10。https://doi.org/10.1038/s41562-017-0189-z

Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., & Keller, M. C.（2019）。多项大型样本研究不支持重度抑郁症的历史候选基因或候选基因与交互作用假设。《美国精神病学杂志》，*176*（5），376–387。https://doi.org/10.1176/appi.ajp.2018.18070881

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R.（2013）。检验效能失败：样本量过小为何会削弱神经科学的可靠性。《自然综述：神经科学》，*14*（5），365–376。https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A.,Chan, T.,Heikensten, E.,Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer,T., Razen, M.,& Wu,H.(2016). 评估

经济学实验室实验的可重复性。《科学》，*351*(6280), 1433–1436。https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., ... Wu, H. (2018). 评估 2010 年至 2015 年间《自然》和《科学》上的社会科学实验的可重复性。《自然·人类行为》，2(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chambers, C. D. (2013). 注册报告：《皮层》的一项全新出版倡议。《皮层》，49(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

CMS合作组。（2012）。利用大型强子对撞机上的CMS实验观测到质量为125 GeV的新玻色子。《物理快报*B*辑》，*716*(1)，30–61。https://doi.org/10.1016/j.physletb.2012.08.021

Cohen, J.（1962）。变态与社会心理学研究的统计功效：综述。《变态与社会心理学杂志》，*65*(3)，145–153。https://doi.org/10.1037/h0045186

Cowan, G., Cranmer, K., Gross, E., & Vitells, O.（2011）。用于检验新物理的基于似然的检验的渐近公式。《欧洲物理学杂志*C*辑》，*71*(2)，文章编号1554。https://doi.org/10.1140/epjc/s10052-011-1554-0

Cummings, J. L., Morstorf, T., & Zhong, K.（2014）。阿尔茨海默病药物开发管线：候选药物寥寥，失败频仍。《阿尔茨海默病研究与治疗》，*6*(4)，文章编号37。https://doi.org/10.1186/alzrt269

Earman, J.（1992年）。非贝叶斯不可？对贝叶斯确证理论的批判性考察。麻省理工学院出版社。

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B.A.（2021）。调查临床前癌症生物学的可重复性。《*eLife*》，*10*，文章e71601。https://doi.org/10.7554/eLife.71601

Gelman, A., & Loken, E.（2014年）。科学中的统计危机。《美国科学家》，*102*（6），460–465。https://doi.org/10.1511/2014.111.460

Gillies，D.（1993年）。*20*世纪的科学哲学：四个核心主题。Blackwell。

Hedges，L. V.（1984）。非随机抽样条件下效应量的估计：删失产生统计学上不显著的均值差异的研究所造成的影响。《教育统计学杂志》，*9*(1)，61–85。https://doi.org/10.2307/1164832

Ioannidis，J. P. A.（2005）。为何大多数已发表的研究发现都是假阳性。《公共科学图书馆医学》，*2*(8)，文章编号 e124。https://doi.org/10.1371/journal.pmed.0020124

Ioannidis，J. P. A.（2008）。为何大多数发现的真实关联都被高估。《流行病学》，*19*(5)，640–648。https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis，J. P. A.（2018）。改革营养流行病学研究的挑战。《美国医学会杂志》，*320*(10)，969–970。https://doi.org/10.1001/jama.2018.11025

Kleijn, B. J. K., 与 van der Vaart, A. W. (2012). 模型设定错误下的伯恩斯坦-冯·米塞斯定理. 《电子统计学杂志》，*6*，354–381. https://doi.org/10.1214/12-EJS675

库恩, T. S. (1962). 《科学革命的结构》. 芝加哥大学出版社.

拉卡托斯, I. (1978年).《科学研究纲领方法论》（J. Worrall 与 G. Currie 编）. 剑桥大学出版社.

Mayo, D. G. (1996). 《错误与实验知识的增长》. 芝加哥大学出版社.

开放科学协作组织。（2015年）。估算心理科学的可重复性。《科学》，*349*(6251)，文章编号 aac4716。https://doi.org/10.1126/science.aac4716

Pe’er, I.、Yelensky, R.、Altshuler, D. 和 Daly, M. J.（2008）。估算几乎所有常见变异的全基因组关联研究的多重检验负担。遗传流行病学，*32*(4)，381–385。https://doi.org/10.1002/gepi.20303

Popper, K. R.（1959年）。科学发现的逻辑。Hutchinson & Co.

Sedlmeier, P., & Gigerenzer, G.（1989年）。统计功效研究会对研究的检验效能产生影响吗？《心理学通报》，*105*(2), 309–316。https://doi.org/10.1037/0033-2909.105.2.309

Simmons,J. P., Nelson, L. D., & Simonsohn, U.（2011年）。假阳性心理学：数据收集和分析中未披露的灵活性，使得任何结果都可以被呈现为显著。《心理科学》，*22*(11), 1359–1366。https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P.（2014年）。P 曲线：破解文件抽屉问题的关键。《实验心理学杂志：总论》，*143*(2), 534–547。https://doi.org/10.1037/a0033242

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. (2004年). 评估阳性报告为假阳性的概率：分子流行病学研究的一种方法。《美国国家癌症研究所杂志》, *96*(6), 434–442. https://doi.org/10.1093/jnci/djh075

White, H. (1982年). 设定错误的模型的最大似然估计。《计量经济学》，*50*(1), 1–25. https://doi.org/10.2307/1912526