arXiv:2302.00774v1 [stat.AP] 1 Feb 2023

# 基于已发表的 $p$ 值估计医学期刊中（随机）临床试验的错误发现风险

预印本

**UlrichSchimmack**[1] **和 František Bartoš**[2,3] 加拿大密西沙加，多伦多大学密西沙加分校心理学系[2] 荷兰阿姆斯特丹，阿姆斯特丹大学心理方法学系[3] 捷克共和国布拉格，捷克科学院计算机科学研究所

通信 如需就本文联系作者，请联系 Ulrich Schimmack，邮箱：ulrich.schimmack@utoronto.c a

这是作者的稿件版本。

## 摘要

「大多数已发表结果都是错误的」这一颇具影响力的论断，引发了人们对科学可信度和诚信的担忧。此后，人们曾多次尝试考察假阳性结果的比例，但这些尝试未能通过实证研究解决这一问题。在此，我们提出一种估计假阳性风险的新方法，并将其应用于顶尖医学期刊中（随机）临床试验的结果。与「大多数已发表结果都是错误的」这一说法相反，我们发现，传统的显著性标准 $\alpha = .05$ 所对应的假阳性风险为 13%。将 $\alpha$ 调整为 .01，可使假阳性风险降至 5% 以下。不过，我们的方法确实提供了明确证据，表明发表偏倚会导致效应量估计偏高。这些结果为评估医学研究的可信度奠定了坚实的实证基础。

## 引言

许多科学领域正面临对已发表结果信心不足的危机 [1]。元科学研究揭示了较低的重复率、较低统计功效的估计值，甚至有科研不端行为的报告 [2]。

基于对真实假设所占比例以及检验这些假设的统计功效的假定，Ioannidis [3] 得出结论：大多数已发表结果都是错误的。事实证明，要检验这一预测并不容易。首先，大规模重复研究 [4, 5, 6] 本质上成本高昂，而且只关注一组有限的预先选定的发现 [7]。其次，对荟萃分析的研究表明，统计功效较低，但很少因此得出零假设为真的结论 [8, 9, 10, 11, 12, 13, 14, 15, 16]（但参见 [17, 18]）。

迄今为止，最有希望的假发现率估算尝试，是 Jager 和 Leek 对医学期刊中 [19] 值的$p$研究。他们从医学期刊的摘要中提取了 5,322 $p$个值，发现统计显著结果中只有 14% 可能是假阳性。这一比例相当可观，但与“大多数已发表结果都是错误的”这一说法并不一致。尽管 Jager 和 Leek 的文章基于实际数据，但它对有关假阳性风险的讨论影响相对有限，可能是因为其研究存在若干局限[20, 21, 22, 23]。

他们的估计方法存在一个问题：难以区分真正的零假设（即效应量恰好为零）与统计功效极低的研究；在后者中，效应量可能非常小，但并非为零。为避免这一问题，我们不估计假阳性的实际比例，而是估计与数据相符的最大比例。我们将这一估计称为错误发现风险（FDR）。为估计 FDR，我们利用

Soric 的 [24] 这一洞见：当检测真实效应的统计功效为 100% 时，错误发现风险达到最大。在这种情况下，假发现率是发现率（即显著结果的百分比）的简单函数。因此，FDR 实证研究面临的主要挑战，是在存在选择偏倚、且这种偏倚会抬高观察到的发现率时，估计发现率。为解决选择偏倚问题，我们开发了一种选择模型，可估计显著性选择之前的发现率。方法部分详细介绍了我们的方法，并将其与 Jager 和 Leek 的 [19]方法进行了比较。

## 方法

为估计假发现率，Jager 和 Leek [19] 开发了一个包含两类研究的模型。一类研究中，零假设成立（$\mathcal{H}_0$）。另一类研究中，零假设不成立，即备择假设成立（$\mathcal{H}_1$）。该模型假设，观测到的显著 $p$值分布是这两类研究的混合分布，并将$p$值的分布建模为截断贝塔分布的混合。该模型的第一个问题是，难以区分 $\mathcal{H}_0$ 成立的研究和 $\mathcal{H}_1$ 成立但统计功效较低的研究。第二个问题是，已发表研究的统计功效存在异质性；在使用截断贝塔分布对 $\mathcal{H}_1$ 下的 $p$值分布进行建模时，可能无法恰当反映这种异质性。为避免第一个问题，我们不估计实际的假发现率，因为这需要明确区分零假设成立与不成立（即区分效应量为零与不为零）。相反，我们旨在估计与数据相符的最大假发现率。我们将这一数值称为错误发现风险。为避免第二个问题，我们采用 z 曲线，将观测到的 $z$统计量分布建模为截断折叠正态分布的混合。该混合模型允许在 $\mathcal{H}_1$ 成立时，研究的统计功效存在异质性 [25, 26]。

### 错误发现风险

为了估计错误发现风险，我们借鉴 Soric 的 [24] 见解：错误发现率的上限受检测真实效应的统计功效所限。当统计功效为 100% 时，所有不显著结果都是通过检验错误假设得出的（$\mathcal{H}_0$）。由于这种情形下不显著 $\mathcal{H}_0$的数量最多，因此显著 $\mathcal{H}_0$ 检验的数量也最多，错误发现率也最高。例如，如果 100 项研究产生了 30 个显著结果，发现率就是 30%。而当发现率为 30% 时，$\alpha = 0$为 .05 时错误发现风险的上限为 ≈0.12。一般来说，错误发现风险是发现率经过简单变换后的结果，例如

$$\text{false discovery risk} = \frac{1 - \text{discovery rate}}{\text{discovery rate}} \times \frac{\alpha}{1 - \alpha}.$$

如果所有实施的假设检验都得到报告，那么只需计算显著结果所占的百分比，就能确定错误发现风险。然而，众所周知，期刊发表统计显著结果的可能性高于发表不显著结果的可能性。这种选择偏倚使期刊中观察到的发现率无法提供有用信息 [26, 25]。因此，任何对错误发现风险进行经验估计的主要挑战，都是将选择偏倚纳入考量。

#### Z-curve 2.0

Brunner 和 Schimmack [25] 开发了一种混合模型，用于估计经过显著性选择后研究的平均功效。Bartoš 和 Schimmack [26] 最近发表了该方法的扩展版本，可根据仅使用统计显著的 $p$ 值拟合模型所得的权重，估计显著性选择前的预期发现率。该选择过程假设研究发表的概率与其统计功效成正比 [25]。大量模拟研究表明，z 曲线能够对预期发现率给出良好的大样本估计 [26]。此外，这些模拟研究还表明，z 曲线得出的置信区间具有稳健性，且覆盖率良好。由于错误发现风险是预期发现率的简单变换，这些置信区间也能为错误发现风险估计提供置信区间。

z 曲线通过估计研究在显著性筛选之前的平均统计功效，得出预期发现率的估计值。z 曲线利用了 $z$ 统计量的分布特性——统计功效相同的研究会产生呈正态分布的 $z$ 统计量，其中心位于真实统计功效经过 a $z$ 转换后的位置（例如，Bartoš 和 Schimmack [26] 中的公式 3）。z 曲线通过在相应的 $\alpha$ 水平处截断来纳入对统计显著性的筛选，这与 Jager 和 Leek 的方法类似。然而，与其他方法不同，z 曲线承认研究的统计功效天然存在异质性（即研究考察的效应不同，样本量也不同）

并通过混合模型估计平均统计功效，

$$f(z;\theta) = \sum_{j=1}^{J} \pi_j f_{j[a,b]}(z;\theta_j),$$

其中， $z$ 对应于统计显著的 $z$ 统计量，这些统计量通过截断折叠正态分布的 $J$ 混合模型进行建模， $f(z;\theta)$，其中 $\theta$ 表示截断折叠正态分布的参数。[1] 有关直观示意，请参见 Bartoš 和 Schimmack [26] 中的图 1。

我们扩展了 z 曲线模型，以纳入经过四舍五入的 $p$ 值，以及以不等式形式报告的 $p$ 值（即 $p < 0.001$）。我们采用了与 Jager 和 Leek 的方法相同的做法：实现区间删失（针对经过四舍五入的 $p$ 值）和右删失（针对以不等式形式报告的 $p$ 值）似然版本的 z 曲线模型，

$$f^{[l,u]}(z;\theta) = \sum_{j=1}^{J} \pi_j f^{[l,u]}_{j[a,b]}(z;\theta_j),$$

其中， $l$ 和 $u$ 对应于截断折叠正态分布混合模型中的下限和上限删失点。例如，以不等式形式报告的 $p$ 值 $p < 0.001$，只有一个下限删失点，即对应于 $p = 0.001$ 的 $z$ 分数；而经过四舍五入的 $p$ 值 $p = 0.02$，下限删失点对应于 $p = 0.025$ 的 $z$ 分数，上限删失点对应于 $p = 0.015$ 的 $z$ 分数。

## 模拟研究

我们进行了模拟研究，在多个方面扩展了 Jager 和 Leek [19] 所做的模拟。我们没有直接模拟 $\mathcal{H}_1$ $p$值（即没有像 Jagger 和 Leek 那样使用右偏贝塔分布）。相反，我们根据双侧 $z$检验模拟了 $\mathcal{H}_1$ $p$值。我们根据 Lamberink 及其同事 [16] 对医学临床试验统计功效的经验估计所建立的分布，为每项模拟的 $z$检验抽取统计功效（唯一需要的参数；排除所有基于含有不显著结果的荟萃分析得出的统计功效估计）。这使我们能够评估各种方法在检验功效存在异质性的情况下，检测与实际医学文献相对应的 $\mathcal{H}_1$ 的表现。为了模拟 $\mathcal{H}_0$ $p$值，我们采用了均匀分布。

我们将真实假发现率从 0 调整到 1，步长为 0.01，并通过改变分别从 $\mathcal{H}_0$ 和 $\mathcal{H}_1$模拟的 $p$值所占比例，模拟了 10,000 个观测到的显著 $p$值。与 Jager 和 Leek[19], 类似，我们设置了四种模拟情景，逐步增加报告不精确的 $p$值所占比例。情景 A 使用精确的 $p$值；情景 B 将 $p$值四舍五入到小数点后三位（低于 0.001 的 $p$值按 0.001 处理）；情景 C 将 20% 的 $p$值四舍五入到小数点后两位（四舍五入为 0 的 $p$值按 0.01 处理）；情景 D 则先将 20% 的 $p$值四舍五入到小数点后两位，再将 20% 的 $p$值截断到最接近的一个上限（0.05、0.01 或 0.001）。

[1] 引入折叠是因为 我们不知道效应的预期方向 。

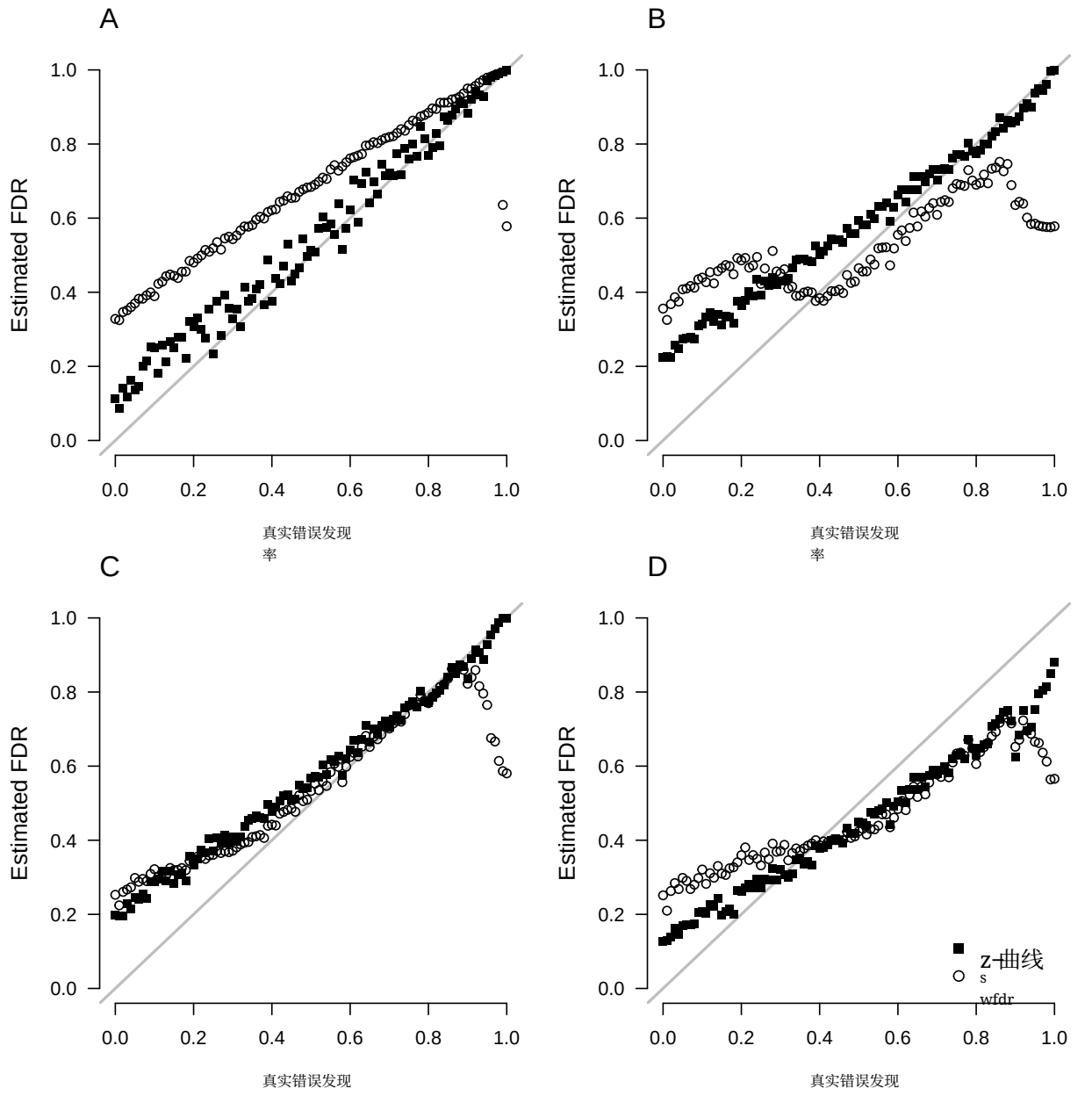


图 1：通过模拟研究比较 z 曲线与 Jager 和 Leek 的假发现风险估计。对于 z 曲线（黑色实心方块）和 Jager 和 Leek 的 swfdr（空心圆），估计假发现率（FDR，纵轴）与真实假发现率（FDR，横轴）的关系，展示了四种模拟情景（各面板）。情景 A 使用精确的 $p$值；情景 B 将 $p$值四舍五入到三位小数；情景 C 将 20% 的 $p$值四舍五入到两位小数；情景 D 将 20% 的 $p$值四舍五入到两位小数，并将 20% 的 $p$值截尾至最接近的显著性水平之一（0.05、0.01 或 0.001）。

图1展示了不同情景（各面板）下，Jager 和 Leek 的 [19] swfdr 方法的真实假发现率（$x$轴）与估计假发现率（$y$轴），以及 z 曲线的假发现风险。当报告精确的 $p$值时（左上角的面板 A），z 曲线能够很好地处理整个假发现率范围内的效力异质性，并准确估计假发现风险。由于假发现风险是对假发现率上限的估计，因此预计其估计值会高于实际假发现率。当真实假设检验的统计功效较低时，尤其容易出现这种差异。在模拟情景中，差异小于 20 个百分点，并且随着真实假发现率增加而减小。尽管 Jager 和 Leek 的 [19]

该方法旨在估计真实的假发现率，但其估计值高于 z 曲线。这是有问题的，因为该方法会夸大真实假发现率的估计值。即使将这些估计值解释为最大估计值，与 z 曲线方法相比，该方法对假发现率实际变化的敏感性也较低。

B 面板显示，当 $p$ 值舍入到小数点后三位时，z 曲线方法会产生相似的结果。然而，Jager 和 Leek 的 [19] 方法会遇到估计问题，尤其是在真实假发现率较低的范围内，因为当前的 swfdr 实现仅允许处理舍入到小数点后两位的情况（我们还尝试将 $p$ 值指定为舍入后的输入；但优化程序出现了多项错误而无法运行）。

C 面板显示，当 20% 的 $p$ 值舍入到小数点后两位时，这两种方法的表现出人意料地相似；但在真实假发现率非常高的情况下，Jager 和 Leek 的 [19] 方法开始低估假发现率。尽管两者表现相似，但由于这两种方法在假发现率较低时都会高估真实假发现率，因此结果必须解释为错误发现风险（最大假发现率）的估计值。
D 面板显示，当 20% 的 $p$ 值处于 .05、.01 或 .001 这一最接近的上限，且未提供有关确切 $p$ 值的明确信息时，两种方法都会出现问题。对真实假发现率超过 40% 的情况进行低估，并非特别严重的问题，因为任何实际假发现率超过 40% 都高得令人无法接受。在这种情形下，解决低估问题的一种可能方法是，在分析中排除被删失的 $p$ 值，前提是删失与 $p$ 值的大小相互独立。

表 1：模拟研究中 z 曲线和 Jager 和 Leek 的错误发现风险估计的均方根误差和偏倚。

| 方法（均方根误差） | A | B | C | D |
|---|---|---|---|---|
| z 曲线 | 0.07 (0.00) | 0.12 (0.01) | 0.10 (0.01) | 0.11 (0.01) |
| swfdr | 0.21 (0.01) | 0.21 (0.01) | 0.13 (0.01) | 0.17 (0.01) |

| 方法（偏差） | A | B | C | D |
|---|---|---|---|---|
| z 曲线 | 0.04 (0.00) | 0.09 (0.01) | 0.07 (0.01) | -0.05 (0.01) |
| swfdr | 0.17 (0.01) | 0.02 (0.02) | 0.04 (0.01) | -0.03 (0.02) |

z 曲线和 Jager 和 Leek 的 swfdr 在四种模拟情景（各列）下对假发现率（FDR）的估计值的均方根误差和偏倚 + 标准误（括号内）情景 A 使用精确的 $p$值；情景 B 将 $p$值四舍五入至三位小数；情景 C 将 20% 的 $p$值四舍五入至两位小数；情景 D 将 20% 的 $p$值四舍五入至两位小数，并将 20% 的 $p$值截断至最接近的显著性水平（0.05、0.01 或 0.001）之一。 [19]

表 1 汇总了各情景下假发现率估计的均方根误差和偏倚，结果显示，z 曲线产生的估计值的均方根误差显著较低。偏倚结果表明，两种方法都倾向于高估真实假发现率。z 曲线出现这种情况是意料之中的，因为它旨在估计最大假发现率。只有当错误发现风险的估计值低于实际假发现率时，才会构成问题。这仅出现在情景 D 中；但如前所示，只有在真实假发现率较高时才会发生低估。

总而言之，我们的模拟结果证实，Jager 和 Leek 的 [19] 方法能够对错误发现风险作出有意义的估计，并且该方法更有可能高估而非低估真实的假发现率。我们的结果还表明，z 曲线优于原始方法；当假发现率低于 40% 时，这些改进能够处理舍入和不精确报告问题。只有当估计值超过 40% 时，才有可能大多数已发表结果都是假阳性结果。

## 在医学期刊中的应用

我们开发了一种改进的摘要抓取算法，用于从可通过 https://pubmed.ncbi.nlm.nih.gov/ 访问的主要医学期刊摘要中提取 $p$ 值。我们的改进算法解决了针对 Jager 和 Leek 的 [19] 原始文章所提出的评论中提到的多个问题。具体而言：（1）按照 [20, 23, 21], 的建议，我们仅从标注为“随机对照试验”或“临床试验”的摘要中提取 $p$ 值；（2）按照 [23], 的建议，我们改进了用于提取 $p$ 值的正则表达式脚本，以涵盖更多可能的表示形式；（3）按照 [23, 22] 的建议，我们从未报告 $p$ 值的摘要中提取置信区间。按照 [20] 的建议，我们还从《公共科学图书馆医学》的摘要中抓取 $p$ 值，以便将假发现率估计与一家选择性较低的期刊进行比较。最后，在所有分析中，我们对抓取到的 $p$ 值随机抽样，每篇摘要仅保留一个 $p$ 值，从而打破了

按照 [20] 的建议，我们打破了估计值之间的相关性。尽管所选方法本身存在其他局限性，但这些改进，加上我们改进后的估计方法，使我们能够检验多位评论者提出的预测：假发现率远高于 14%。

我们于2021年7月执行了抓取流程，抓取了自2000年以来发表的摘要（抓取数据汇总见表2）。可在 https://tinyurl.com/zcurve-FDR 查看各篇摘要及抓取值的交互式可视化。

表 2：抓取数据汇总

| 期刊 | 摘要 | 临床试验或随机对照试验 | 可自动抓取 | $p$-值 |
|---|---|---|---|---|
| 《柳叶刀》 | 7638 | 2099 | 1556 | 5403 |
| BMJ | 4480 | 939 | 643 | 1625 |
| NEJM | 4912 | 2767 | 2304 | 7137 |
| JAMA | 5491 | 1452 | 1274 | 4484 |
| PLoS | 2944 | 346 | 279 | 1102 |
| 总计 | 25465 | 7603 | 6056 | 19751 |

“摘要”列包含访问的摘要总数，“临床试验或随机对照试验”列包含标记为“临床试验”或“随机对照试验”的摘要数量，“可自动抓取”列包含至少含有一个可自动抓取的 $p$值或置信区间的其他摘要数量，“$p$值”列包含提取的 $p$值总数。PLoS 指《公共科学图书馆医学》，NEJM 指《新英格兰医学杂志》。

图 2 展示了基于从临床试验和随机对照试验摘要中抓取的数据，z 曲线和 Jager 和 Leek 的 [19] 方法所估计的错误发现风险，并按期刊以及文章发表时间是否在2010年（含）之前（左）或之后（右）进一步划分。可以看到，与模拟结果一致，Jager 和 Leek 的 [19] 方法得出的假发现率估计略高。此外，z 曲线得出的自助法置信区间明显更宽，这表明 Jager 和 Leek 报告的置信区间 [19] （± 1个百分点）过窄。

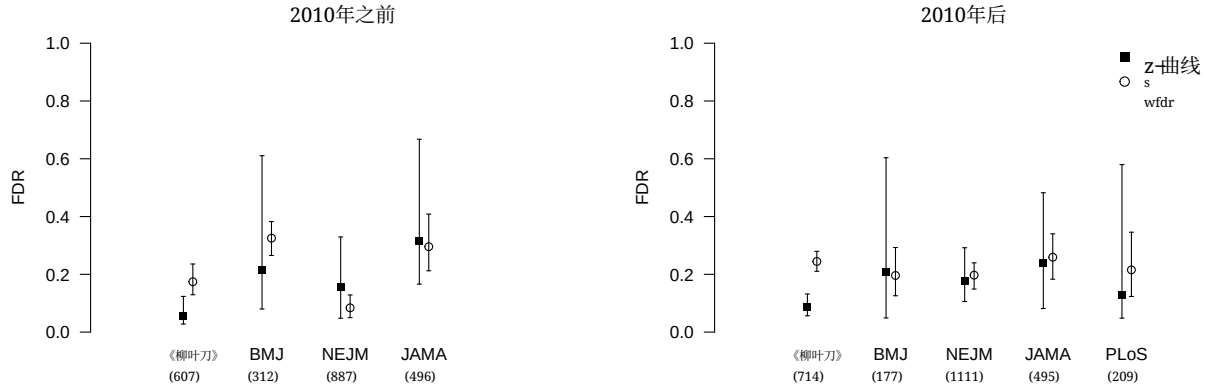


图2：各十年期和期刊中的错误发现风险估计。基于按期刊划分的临床试验和随机对照试验，以及文章是在2010年之前（含2010年）发表（左侧）还是在2010年之后发表（右侧），采用 z 曲线（黑色实心方块）和 Jager 和 Leek 的 [19] swfdr（空心圆）估计假发现率（FDR，纵轴）；横轴为期刊，并在括号中标注独立文章的信息。

比较基于 2010 年及以前的数据和 2010 年之后的数据所得的错误发现估计，可以发现置信区间存在重叠，这表明假发现率没有变化。分别基于临床试验和随机对照试验进行的分析也未显示显著差异（见图 3）。因此，为了减少假发现率估计的不确定性，我们不考虑发表年份，估算各期刊的假发现率。基于 z 曲线和 Jager 和 Leek 的 [19] 方法得出的假发现率估计汇总于表 3。我们发现，所有假发现率估计均落在 0.05 至 0.30 的区间内。最后，将各期刊的数据进一步汇总后，基于 z 曲线得出的假发现率估计为 0.13，95% [0.08，0.21]；基于 Jager 和 Leek 的方法得出的估计为 0.19，95 %

[0.17，0.20] 。这一发现表明，Jager 和 Leek 的 [19] 提取方法略微低估了假发现率，而他们的模型则高估了假发现率。最后，我们基于 z 曲线得出的改进版错误发现风险估计与原始假发现率估计非常接近。

表 3：综合及各期刊的假发现风险估计。

| 期刊 | z 曲线 | swfdr |
|---|---|---|
| 《柳叶刀》 | 0.07 [0.05，0.11] | 0.21 [0.18，0.24] |
| BMJ | 0.23 [0.08，0.54] | 0.28 [0.23，0.32] |
| NEJM | 0.18 [0.10，0.30] | 0.14 [0.11，0.18] |
| JAMA | 0.29 [0.15，0.55] | 0.26 [0.21，0.31] |
| PLoS | 0.12 [0.05，0.54] | 0.21 [0.13，0.32] |
| 合并 | 0.13 [0.08，0.21] | 0.19 [0.17，0.20] |

基于自2000年以来发表的临床试验和随机对照试验，使用 z 曲线和 Jager 和 Leek 的[19] swfdr 方法，估算各分析期刊及合并数据集的错误发现风险（FDR）及其 95% 置信区间。

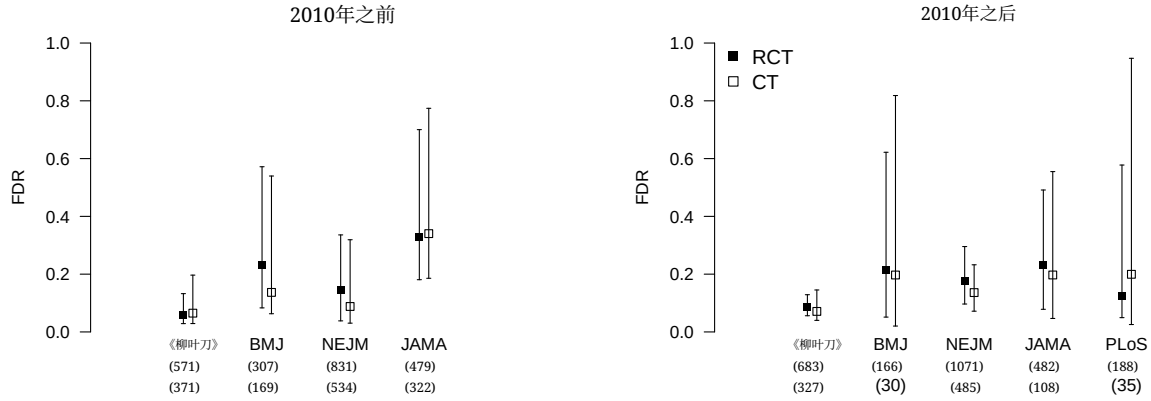


图 3：不同十年期、期刊和研究类型的错误发现风险估计。使用 z 曲线估算随机对照试验（RCT，黑色实心方块）和临床试验（CT，空心方块）的假发现率（FDR，纵轴），并按期刊划分（横轴，括号中为独立文章数量；第一行为 RCT，第二行为 CT），同时区分文章发表于 2010 年（含）之前（左侧）还是 2010 年之后（右侧）。

## 其他 z 曲线结果

到目前为止，我们仅使用预期发现率来估计错误发现风险，但预期发现率本身也能提供有价值的信息。Ioannidis 对临床试验假发现率的 [3]预测基于两种情景。一种情景假设统计功效较高（80%），且真实假设与错误假设的比例为 1:1。这种情景意味着真实发现率为 .5*.8 + .5*.05 = 42.5%。第二种情景假设统计功效为 20%，且每 5 个错误假设对应 1 个真实假设。这种情景意味着真实发现率为 .17*.20 + .83*.05 = 7.5%。不同情景中的偏倚程度也有所不同，偏倚会抬高观察到的发现率。不过，z 曲线会校正这种偏倚。因此，将 z 曲线对发现率的估计与 Ioannidis 对临床试验发现率的假设进行比较颇有意义。Z 曲线对预期发现率的估计为 30%，95% 置信区间为 20% 至 41%。这一发现表明，我们样本中的大多数临床试验已发表结果来自统计功效充分、且很可能检验真实假设的研究。这一结论也与我们的错误发现风险估计相符：该估计与 Ioannidis 对高质量临床试验的预测相近，假发现率 = 15%。因此，对 Ioannidis 的主张与我们的结果之间差异的一种简单解释是，Ioannidis 低估了文献中统计功效充分的研究所占比例。

预期发现率还提供了有关医学期刊中选择偏倚程度的重要信息。预期发现率仅为 30%，而观察到的发现率（即摘要中显著结果所占的比例）却超过其两倍（69.7%）。这一差异在图 4 中清晰可见。观察到的非显著 z 分数直方图与预测分布（蓝色曲线）并不吻合。这表明存在相当严重的选择偏倚

这意味着报告的效应量因选择偏倚而被夸大。因此，在通过功效分析规划样本量时，后续研究需要对效应量进行调整。此外，荟萃分析需要校正选择偏倚，以获得无偏的平均效应量估计 [27]。

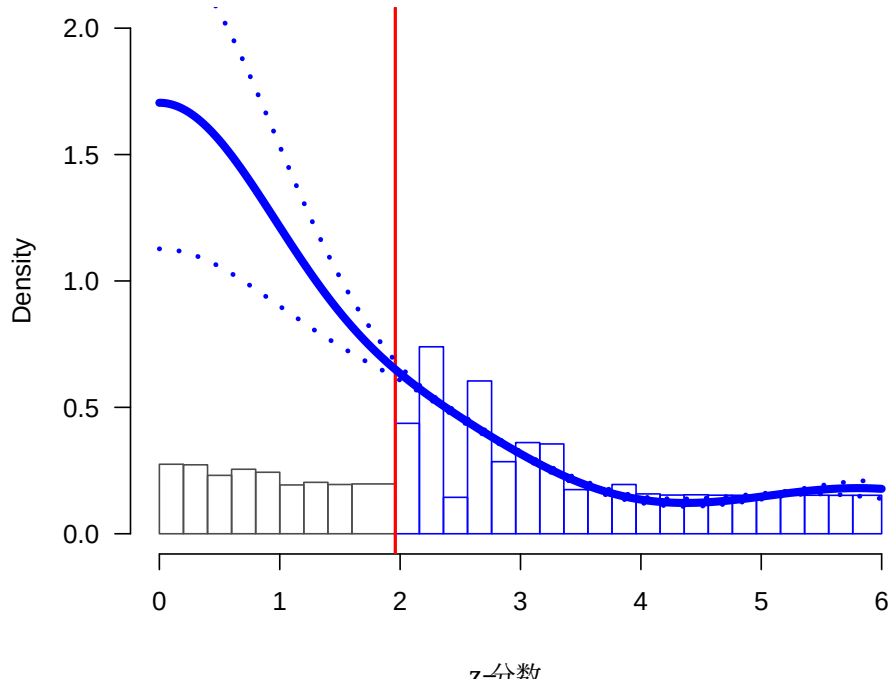


图 4：合并数据集的 Z 曲线。所有显著 $p$ 值转换为 $z$ 统计量后的分布，以及拟合的 Z 曲线密度（蓝色实线）和 95% 逐点置信带（蓝色虚线）。

Z 曲线还提供了医学摘要中统计显著结果预期重复率的信息，该重复率定义为：在样本量和显著性标准相同的近似重复研究中获得统计显著结果的概率（其他定义见例如 [28, 29, 30] ）。预期重复率并非错误发现风险的简单补数，因为真实发现可能因重复研究的统计功效较低而无法重复，而错误发现也可能以与 $\alpha$ 相对应的概率“错误地”重复。预期重复率为 65%，置信区间为 61% 至 69%。重要的是要认识到，这一预期重复率包含真实假设和错误假设的重复结果。虽然实际假阳性率未知，但我们可以使用 Z 曲线估计值，假设显著结果中有 14% 是假阳性。在此假设下，对真实假设进行重复研究的统计功效为 75%。相比之下，再次获得假阳性结果的概率仅为 0.7%。因此，即使样本量相同，实际重复研究也能成功降低假阳性风险。然而，假阴性风险仍然很高：重复研究中的不显著结果有 44% 检验的是真实假设。

此外，z 曲线对重复率的估计高于实际重复率 [26]。原因之一可能是，精确的重复研究并不可能实现，而且人群变化会因选择偏倚和回归均值而导致统计功效降低。在最坏的情况下，实际重复率可能低至预期发现率。因此，我们的结果预测，医学领域实际重复研究的成功率将在 30% 至 65% 之间。为了确保有较高概率重复出真实结果，因此有必要增加重复研究的样本量。如果无法做到这一点，就应谨慎解读未成功的重复结果，并开展进一步研究。单次重复研究失败并不能得出定论。

Ioannidis [23] 提出，14% 的假发现率是一个理想目标。然而，许多研究人员错误地认为，$\alpha$= .05 能确保错误发现风险只有 5%。此外，$\alpha$ = .05 长期沿用，表明许多研究人员认为 5% 的错误率可以接受。我们证明，$\alpha$ = .05 会导致 14% 的假发现率。要将假发现率降至 5% 或以下，就必须降低 $\alpha$。Z 曲线分析有助于找到符合这一标准的 $\alpha$ 水平。将 $\alpha$ =.01 作为标准时，预期发现率会降至 20%，错误发现风险则会降至 4%。根据这些结果，可以使用 $\alpha$ = .01 作为拒绝零假设的标准，同时将假阳性风险维持在 5% 以下。

虽然期刊可能不愿意规定较低的 $\alpha$水平，但读者可以采用这一标准，将接受错误假设的风险维持在合理的低水平。

## 讨论

与许多其他人类活动一样，科学依赖信任。Ioannidis（2005）的文章指出，人们没有理由信任科学，而且读者在科学期刊上遇到错误主张的可能性高于正确主张。为回应这些担忧，元科学这一新领域应运而生，旨在考察已发表结果的可信度。我们的文章在这些努力的基础上，提出一种改进的方法，用于估计医学期刊中的错误发现风险，重点关注临床试验。与 Ioannidis 所认为的这类结果大多可能是假阳性不同，我们发现，按照传统的统计显著性标准（$\alpha$ = .05），错误发现风险为 14%；按照较低的阈值 $\alpha$ = .01，则为 4%。因此，我们的结果在一定程度上令人放心，表明医学研究比 Ioannidis 预想的更为稳健。

与此同时，我们的结果也揭示了一些期刊编辑可以轻易解决的问题。最大的问题仍然是偏袒统计显著结果的发表偏倚。显著性选择会造成许多不良后果。尽管医学界已通过要求临床试验进行预注册来应对这一问题，但我们的结果表明，显著性选择仍是医学研究中普遍存在的问题。z 曲线的一项创新贡献是能够量化选择偏倚的程度。期刊编辑可以利用这一工具追踪选择偏倚并制定政策加以减少。这可以通过多种方式实现。首先，编辑可以发表更多设计严谨、结果不显著的重复研究。其次，编辑可以更仔细地核查研究人员是否遵循了预注册方案。最后，编辑可以优先考虑样本量较大的研究。

尽管我们对 Jager 和 Leel 的研究作出了改进，但我们的研究仍有一些局限性，可在未来研究中加以解决。一个值得关注的问题是，摘要中的结果可能无法代表结果部分中的其他结果。另一个问题是，我们选择的期刊不具代表性。这些担忧是合理的，也意味着我们的结论仅适用于我们考察的特定期刊。将结论推广到其他科学学科或其他医学期刊是不妥的。我们表明，z 曲线是一种有用的工具，可用于估计假阳性风险，并调整 $\alpha$ 以将假阳性风险控制在理想水平。该工具可用于估计其他学科和研究领域的错误发现风险，并设定与这些领域发现率相符的 $\alpha$ 水平。我们认为，对发现率、假阳性风险、选择偏倚和重复率进行经验估计，有助于制定科学政策，确保已发表结果为科学进步奠定坚实基础。

## 数据和代码的可用性

补充材料（包括用于复现模拟、数据抓取和分析的数据及 R 脚本）可从 https://osf.io/y3gae/ 获取。zcurve R 包可从 https://cran. r-project.org/package=zcurve 获取。

## 参考文献

[1] Baker M. 可重复性危机。《自然》2016; 533(26): 353–66.[2] Fanelli D. 观点：科学真的正面临可重复性危机吗？我们需要它吗？ 美国国家科学院院刊2018; 115(11): 2628–2631.[3] Ioannidis JP. 为什么大多数已发表研究结果都是错误的。《公共科学图书馆医学》2005; 2(8): e124.[4] 开放科学合作组织。心理科学可重复性的估计。《科学》2015; 349(6251).[5] Camerer CF, Dreber A, Forsell E, et al. 评估经济学实验室实验的可重复性。 《科学》2016; 351(6280): 1433–1436.[6] Camerer CF, Dreber A, Holzmeister F, et al. 评估2010年至2015年间发表于《自然》和《科学》的社会科学实验的可重复性。 《自然·人类行为》2018; 2(9): 637–644.[7] Coyne JC. 重复研究倡议无法挽救心理学的可信度。 *BMC*心理学2016; 4(1): 1–11.[8] Fanelli D, Costas R, Ioannidis JP. 对科学偏倚的元评估。 美国国家科学院院刊2017; 114(14): 3714–3719.[9] Wooditch A, Fisher R, Wu X, Johnson NJ. P值问题？对犯罪学中证据价值的考察。《定量犯罪学杂志》2020; 36(2): 305–328。

[10] Barnes J, TenEyck MF, Pratt TC, Cullen FT. 犯罪学中的证据有多强？我们是否应该担心信心危机即将到来。《司法季刊》2020; 37(3): 383–409.[11] Nuijten MB, Assen vMA, Augusteijn HE, Crompvoets EA, Wicherts JM. 智力研究中的效应量、统计功效和偏倚：一项元元分析。《智力杂志》2020; 8(4): 36. [12] Stanley TD, Carter EC, Doucouliagos H. 荟萃分析揭示了心理学研究的可重复性。《心理学通报》2018; 144(12): 1325.[13] Ioannidis JPA, Stanley TD, Doucouliagos H. 偏倚在经济学研究中的影响力。《经济学杂志》2017; 127(605): F236-F265. doi: 10.1111/ecoj.12461[14] Van Aert RC, Wicherts JM, Van Assen MA. 对心理学和医学荟萃分析中的发表偏倚进行考察：一项元元分析。《公共科学图书馆·综合》2019; 14(4): e0215052.[15] Mathur MB, VanderWeele TJ. 估计同行评审研究荟萃分析中的发表偏倚：跨学科和期刊层级的元元分析。《研究综合方法》2021; 12(2): 176–191.[16] Lamberink HJ, Otte WM, Sinke MR, et al. 临床试验的统计功效有所提升，而效应量保持稳定：对1975年至2014年间136,212项临床试验的实证分析。《临床流行病学杂志》2018; 102: 123–128.[17] Bartoš F, Gronau QF, Timmers B, Otte WM, Ly A, Wagenmakers EJ. 医学中的贝叶斯模型平均元分析。《医学统计学》2021; 40(30): 6743-6761.[18] Bartoš F, Maier M, Wagenmakers EJ, et al. 发表选择偏倚在医学、环境科学、心理学和经济学荟萃分析中的影响。预印本见：https://arxiv.org/abs/2208.12334；2022[19] Jager LR, Leek JT. 科学领域假发现率的估计及其在顶尖医学文献中的应用。《生物统计学》2014; 15(1): 1–12.[20] Goodman SN. 讨论：科学领域假发现率的估计及其在顶尖医学文献中的应用。《生物统计学》2014; 15(1): 13–16. [21] Gelman A, O'Rourke K. 根据已发表p值的分布推断科学真理的困难。《生物统计学》2014; 15(1): 18–23.[22] Benjamini Y, Hechtlinger Y. 讨论：Jager 和 Leek 关于科学领域假发现率估计及其在顶尖医学期刊中应用的研究。《生物统计学》2014; 15(1): 13–16.[23] Ioannidis JP. 讨论：为什么"科学领域假发现率的估计及其在顶尖医学文献中的应用"是错误的。《生物统计学》2014; 15(1): 28–36.[24] Sori´c B. 统计"发现"与效应量估计。《美国统计学会杂志》1989; 84(406): 608–610.[25] Brunner J, Schimmack U. 在存在异质性和显著性选择的条件下估计总体平均统计功效。《元心理学》2020; 4.[26] Bartoš F, Schimmack U. Z-curve 2.0：估计重复率和发现率。《元心理学》2021.[27] Bartoš F, Maier M, Wagenmakers EJ, Doucouliagos H, Stanley TD. 稳健贝叶斯元分析：对互补的发表偏倚校正方法进行模型平均。《研究综合方法》2022.[28] Held L, Micheloud C, Pawel S. 基于相对效应量评估重复成功。《应用统计年刊》2022; 16(2): 706 – 720.[29] Pawel S, Held L. 用于评估重复成功的怀疑性贝叶斯因子。英国皇家统计学会杂志：B辑（统计方法学）2022; 84(3): 879–911.[30] Ly A, Etz A, Marsman M, Wagenmakers EJ. 基于证据更新的重复贝叶斯因子。《行为研究方法》2019; 51(6): 2498–2508。