arXiv:2602.06379v1 [stat.ME] 6 Feb 2026

# 自适应临床试验的 e 值：实践中的任意时点有效监测



**Alexandra Sokolova**
Knight 癌症研究所
俄勒冈健康与科学大学
sokolova@ohsu.edu

**Vadim Sokolov**系统工程与运筹学系 乔治·梅森大学
vsokolov@gmu.edu

2026年2月9日

**摘要**

适应性临床试验日益依赖中期分析、灵活停止和依赖数据的设计修改。这些特点可以提升伦理性和效率，但若反复查看固定时程检验统计量，或在适应性调整后重复使用这些统计量，就会使统计保证变得复杂。E值和e过程提供了一种实用的替代方法：它们可产生任意时点有效检验和置信序列，即使采用可选停止或可选延续，也能保持有效性，且无需预先规定监测计划。

本文是面向实践者的方法指南。我们介绍适用于双臂随机对照试验的e过程投注鞅构造，说明e值如何自然地处理复合原假设并支持无效性监测，并就何时适合采用e值、何时优先选择成熟的替代方法，以及如何将e值监测与组序贯和贝叶斯自适应工作流程相结合提供指导。

一项数值研究在一项二元终点双臂试验中比较了五种监测规则——频率学派、贝叶斯和E值方法的朴素版与校准版。频繁进行中期观察时，朴素重复检验和朴素后验阈值会显著抬高第一类错误率。在有效方法中，校准的组序贯规则检验效能最高；E值规则以适中的检验效能提供稳健的任意时刻有效控制；校准的贝叶斯规则则最为保守。

扩展模拟表明，组序贯方法与e值方法之间的功效差距取决于监测计划，并且在连续监测下会反转。该方法学涵盖无效性监测、平台试验多重性控制，以及结合e值与成熟方法的混合策略，已在开源R软件包 `evalinger` 中实现，并置于美国食品药品监督管理局2026年1月关于贝叶斯方法学的指南草案所构建的监管框架之下。

***K*关键词** 适应性临床试验 •期中监测 •可选停止 • e值 • e过程 • 组序贯方法 • α消耗 • 贝叶斯自适应设计 • 安全检验 • 投注鞅 • R软件包

# 1 引言

在实践中，即使统计分析计划写得仿佛推断只会在固定的最终样本量下进行一次，临床试验也具有序贯性。入组和结局累积随着日历时间推进；数据监测委员会会审查不断累积的安全性和疗效信息；操作限制和新出现的外部证据可能促使修订方案。实际问题不在于是否会做出期中决策，而在于如何做出这些决策，同时不牺牲证据的可靠性。

量化一个已实现序列“看起来有多随机”的需求，早已出现在远离临床试验的领域。柯尔莫哥洛夫关于测量有限序列随机性的研究，及其在遗传学和信息论等领域的应用，探讨了这样一些情境：观测结果依次到来，而一旦看到显著模式，人们便总是忍不住停止观察或重新评估（Ramdas 等，2023年）。Ville 于1939年对随机性的鞅刻画比 Wald 序贯概率比检验早了数年，并确立了一个关键不等式：如果投注者的资本是非负上鞅，那么它曾经超过 $1/\alpha$ 的概率至多为 $\alpha$（Ville，1939年；Wald，1947年）。后来，Shafer 和 Vovk 在博弈论概率框架下将这一思想重新表述为明确的“押注式检验”框架，指出它为序贯推断奠定了基础，无需固定样本量或预先设定停止规则（Shafer 和 Vovk，2019年；Shafer，2021年）。现代 e 值正是此类押注的收益，而 e过程则是针对原假设进行序贯押注的策略所产生的财富轨迹（Polson、Sokolov 和 Zantedeschi，2026年）。检验中的投注隐喻与试验设计也有相通之处：Berry 将临床试验置于老虎机问题框架下的决策分析思路同样涉及序贯押注——在这里，对每位患者的“押注”就是治疗分配，而收益则是患者的结局（D. A. Berry，2025）。Thompson 于1933年提出按照某组当前为最佳的信念概率将患者随机分配到各组（Thompson 1933 年的提议），这是早期的自适应分配策略，早于 Hill 的固定随机化和 Wald 的序贯检验；它与 e 值有一个共同特点：每一步的分配（或押注）都是累积证据的可预测函数。对于临床试验而言，基于押注的检验与基于押注的分配汇合而成一个清晰的框架：每当患者结局出现时，就针对原假设下注；无论观察了多少次，累积证据都仍然可以解释。

在当代试验中，经典频率学派推断通过组序贯设计和 α 消耗函数实现期中监测，但代价是必须预先设定监测框架（包括一组中期观察，或一个消耗函数及信息时间的定义），并且正式的第一类错误保证仅适用于该框架（Pocock，1977年；O’Brien 和 Fleming，1979年；Lan 和 DeMets，1983年；Jennison 和 Turnbull，2000年）。贝叶斯自适应设计则借助后验分布和预测概率，采用连续学习与决策，并通过模拟评估操作特征，利用保守的决策阈值加以控制（Spiegelhalter、Abrams 和 Myles，2004年；S. M. Berry 等，2010年；D. A. Berry，2004年、2025年；美国食品药品监督管理局，2025年）。Meurer、Lewis 和 Berry（Meurer、Lewis 和 Berry，2012年）认为，自适应设计能够在一定程度上纠正“治疗性误解”——即一种常见的误解，认为参加试验就能保证患者获得个体化优化治疗——因为随着证据不断积累，自适应设计会让更多患者接受有效治疗。

Berry 的研究为理解“贝叶斯监测”在实际操作中的含义提供了面向实践者的参照（D. A. Berry 1989, 2025）。在 Berry 的决策分析框架中，临床试验本质上是一个老虎机问题：治疗试验入组患者的重要性，不亚于治疗试验结束后将就诊的患者；因此，设计应力求在获益（为当前患者提供有效治疗）与学习（获取能够改善未来治疗决策的信息）之间取得平衡（D. A. Berry 2004; D. A. Berry and Fristedt 1985）。这一思路自然引出通过 Thompson 抽样（Thompson 1933）进行反应自适应随机化：下一位患者被分配至某组的概率，与该组当前为最佳治疗组的后验概率成正比。这一策略比现代 e值早了几十年，但与 e值有着共同的投注理论渊源。在 Berry 的框架中，实际决策所依据的量并非优效性的原始后验概率，而是试验成功的贝叶斯预测概率：在当前后验分布下进行积分，计算若继续入组，试验最终得出积极结论的概率。该指标决定了 CALGB 49907 的停止规则：当获得有意义答案的预测概率超过 80% 时，试验在入组 633 名患者后停止（计划上限为 1800 名）（Muss et al. 2009）；ThermoCool AF 器械试验在首次中期分析时，预测概率达到 99.9%，于是停止（D. A. Berry 2025）；而受老虎机问题启发的平台试验 I-SPY 2 则通过每月评估预测概率，决定 23 种在研疗法和 10 种分子特征对应治疗组的毕业、无效性判定及反应自适应随机化（Barker et al. 2009; Park et al. 2016; D. A. Berry 2025）。需要特别指出的是，贝叶斯试验学家对此十分了解，但实践中有时容易忽视：在可选停止的情况下，基于后验概率或预测概率阈值的规则并不会自动控制频率学派第一类错误；只要设计规格要求提供此类保证，就必须进行校准，通常是在设计阶段通过模拟来完成（U.S. Food and Drug Administration 2025, 2026）。

e值——在原假设下期望不超过一的非负统计量——及其序贯对应物 e过程，为证据和错误控制提供了互补视角。其定义性质保证检验

在可选停止和可选延续下均有效，且无需固定样本量或预先规定的监测计划（Grünwald、Heide 和 Koolen 2023；Ramdas 等，2023年；Shafer，2021年）。这种灵活性简化了监测：同一阈值 $E_t \geq 1/\alpha$ 可在任意查看数据的情况下控制第一类错误。与此同时，e值也带来了新的设计自由度：构造有检验效能的e过程取决于投注策略、备择模型或类先验调节参数的选择，而设定错误可能降低检验效能或增加期望样本量（Grünwald 等人 2021；Martin 2025）。此外，监管生态系统已深度习惯于 $p$值、置信区间以及确证性场景中的家族错误率控制；将 e值转换为报告惯例需要谨慎说明，在某些场景下还需要额外的方法学。

本文有四个目标。第一，实用且以临床为导向地介绍 e 值和 e过程，将其作为中期监测和自适应试验实施的工具，并明确给出适用于二分类和生存终点双臂试验的构造方法。第二，为从业者提供指南，说明何时应使用 e 值、何时优先采用成熟的备择方法，以及无效性、多重性和监管要求如何与 E值方法相互影响。第三，通过一个双臂随机试验中的五种监测规则比较，进行数值演示：朴素的重复 $p$值、类O’Brien–Fleming群序贯界值、投注鞅E过程，以及朴素和经模拟校准的贝叶斯后验阈值。在有效方法中，经校准的组序贯规则检验效能为 86.1%，e过程为 72.3%，校准的贝叶斯规则为 68.8%；而在进行 20 次中期分析时，朴素方法会使第一类错误率膨胀至 13–15 %。第四，本文提供 `evalinger` R软件包及配套的交互式网络应用，实现完整的方法体系，包括设计校准、实时监测、置信序列、无效性分析、平台试验多重性控制，以及与组序贯方法和贝叶斯方法的直接比较，使从业者能够在 2026年FDA贝叶斯指导草案（美国食品药品监督管理局 2026年）的监管框架内应用 e 值监测。

为明确本文的研究范围：第4–5节回顾已有理论（组序贯方法、e值、e过程、Ville不等式、投注鞅构造以及 GROW 最优下注比例），并参考 Ramdas 等人（2023）、Grünwald、Heide 和 Koolen（2023）以及 Shafer（2021）。本文的原创贡献包括：(i) 五种方法的数值比较，其中既包括朴素和校准的贝叶斯规则，也包括组序贯方法和 e 值监测（第7节）；(ii) 扩展模拟，量化检验效能差距对监测计划、设计参数和试验规模的依赖性，并展示连续监测下的反转现象（第7.5–7.8节）；(iii) 通过置信序列和倒数 e 过程构建无效性监测框架，并进行蒙特卡洛评估（第6.3节和第7.7节）；(iv) 在 Novick（1965）的真实数据示例中，对经典方法、贝叶斯方法和 e 值方法进行三方比较（第 8 节）；(v) 提供从业者指南和监管分析，阐明 e 值在 2026 年 FDA 贝叶斯指导草案中的定位（第2–3节和第9节）；以及 (vi) `evalinger` R软件包和交互式网络应用。

## 2 从业者指南：何时在临床试验中使用e值

当计划外或频繁的数据窥视在操作上很可能发生，或可选延续是现实可能时，e值最为有用。这包括具有不规则期中查看的情形（例如，数据安全监察委员会会议按日历时间而非固定信息分数安排）、快速信号检测很重要的连续安全性监测、涉及样本量重新估计、富集、试验组剔除或增加或反应自适应随机化的自适应方案，以及跨子研究或序贯实验的项目层面的证据累积，其中是否开展后续研究的决定取决于较早的结局（Grünwald、Heide 和 Koolen 2023年）。REMAP-CAP/COVID平台试验说明了这一点：该试验被设计为针对社区获得性肺炎的永久性自适应平台，在世界卫生组织大流行病声明后两天内转向开始入组COVID-19患者，最终在15个国家的66种干预措施中对超过24,000名患者进行了随机分配（Angus 等人 2020年；D. A. Berry，2025）。在这种情形下，分析时间表由不断演变的大流行病而非预先规定的统计计划决定，任意时刻有效推断本会是该试验证据框架的自然选择。

组序贯方法仍是许多确证性场景的默认选择，因为监管机构对其十分熟悉，它针对常见的检验统计量进行了优化，并且在监测计划固定时，在相同第一类错误率下比e值具有更高的检验效能。当中期观察次数固定且预先设定（或由商定的消耗函数规定）、终点属于标准类型，且主要目标是在计划的信息时间达到较高的检验效能时，α 消耗和组序贯边界提供了一种成熟且广受认可的解决方案。在这类场景中，e值可作为辅助证据来源，但并不一定优于已经校准良好的组序贯设计。第10 节的数值研究对此进行了量化：在一项有20次等间隔查看的双臂二分类试验中，校准的成组序贯边界达到86.1%的检验效能，而投注e过程达到72.3%，两者相差约14个百分点，体现了实现任意时刻有效性所付出的代价。

贝叶斯自适应设计在决策问题明确涉及成功概率、效用，或跨分层或试验组学习时表现出色（D. A. Berry，2025年）。在这类情况下，后验分布和预测概率是自然的操作

工具，而 Berry 的决策分析框架为设计、监测和适应提供了成熟的工具箱。然而，正如我们的数值研究所证实的，在固定水平下进行后验阈值停止（例如，Pr($p_T > p_C$ | data) >0.975）可能会在频繁进行中期分析时大幅提高第一类错误率。在我们的例子中，进行 20 次观察时，朴素的后验阈值规则的第一类错误率约为 13.5%。要在 20 次观察时达到 2.5% 的第一类错误率，校准的贝叶斯规则要求后验阈值约为 0.998，这比朴素的 0.975 保守得多。该校准阈值的检验效能为 68.8%（在此设计中比投注e过程更保守），且在备择假设下可稍早停止（每组平均 134 名患者，而校准的组序贯方法和 E 值规则均约为 140 名）。

2.1 无效性监测与贝叶斯工作流的整合 **无效性监测与贝叶斯工作流的整合**

无效性——即由于即使继续入组，治疗也不太可能显示出有意义的获益，因而决定停止试验——在自适应试验中与疗效监测同等重要。e过程通过两种互补机制支持无效性监测。首先，可在每次中期观察时计算治疗效应的置信序列（见第 7节），当置信序列的上界低于最小临床重要差异时，即可宣布无效。按照所选错误水平进行解读，这提供了一个随时有效的信号，表明截至目前累积的数据不支持具有临床意义的效应。其次，可以构建一个“倒数”e过程，用于检验反向假设（治疗组至少以具有临床意义的幅度不劣于对照组）；如果该倒数 e 过程增长到较大数值，则有充分证据反对存在具有临床意义的治疗效应。在实践中，置信序列方法对于与数据安全监察委员会的沟通更加透明，因为它直接传达了可能效应量的范围；而倒数 e 过程则提供了更正式的无效性检验，可整合到多重性调整决策框架中。这两种机制均已在 `evalinger` （`futility_cs()` 和 `futility_eprocess()`）中实现，而 Web 应用程序中的监测仪表板提供实时置信序列可视化，以支持实时无效性评估。

D. A. Berry（2025）明确将临床试验设计界定为不确定性条件下的决策：基于不断累积的数据进行调整是一种特性而非违规，而预测概率则是用于继续/终止决策以及样本量决策的自然量。Berry 的愿景是，驱动各项调整的“观察者”应当是一台配备前瞻性算法的计算机，由该算法规定在试验开始之前即已确定的调整方案。这一原则——算法化、前瞻性的调整——与 e过程的要求直接契合，即投注比例须为既往数据的可预测函数。从这个视角看，e值是对贝叶斯决策规则的补充。当试验使用预测概率阈值来采取操作性行动时——例如 AWARD-5 无缝衔接 II 期和 III 期，借助每两周一次的贝叶斯分析，从七个候选剂量中选出度拉糖肽的两个剂量，且无需暂停入组（Geiger 等，2012；D. A. Berry，2025）——可以并行运行 e过程，将其作为可审计的、任意时刻有效的证据账本，用于 I 类错误控制。这并不取代贝叶斯决策规则；相反，它为中期分析中的疗效主张提供了一个保守且透明的频率学派安全层。这种组合对于 GBM AGILE 之类的无缝衔接分期设计十分自然：在该设计中，试验组第 1 阶段（反应自适应）与第 2 阶段（确证性扩展）的患者被合并，用于单一注册分析（Alexander 等，2018；D. A. Berry，2025）：一个在两个阶段持续累积证据的 e过程能够提供任意时刻有效的错误控制，而无需各阶段分别制定统计分析计划。当调整逻辑复杂且高度依赖模拟校准时，e值还可以简化监测治理，因为它们提供单一的时间一致阈值，该阈值在观察次数安排发生变化时仍然有效，从而减轻临时修改界值的压力。

### 2.2 软件、监管环境与常见陷阱

临床试验中用于构造 e 值的软件日趋成熟。 `safestats` R软件包（Schure和Ly，2022年）实现了适用于常见终点的安全检验，包括用于生存数据的安全对数秩检验、安全 $t$检验，以及比例安全检验，并提供校准工具，以评估不同设计备择下的检验效能和期望样本量。更广泛的 e 值文献也有 R 和 Python 研究软件的支持，不过，与组序贯方法相比，达到生产级别的临床试验软件仍不够成熟（例如， `gsDesign`、 `rpact`）。从监管角度来看，美国食品药品监督管理局关于自适应设计的指南强调，任何自适应方法都必须前瞻性地控制第一类错误并维持试验完整性（美国食品药品监督管理局，2019年）。e值通过其构造方式满足这一要求（Ville不等式提供解析保证），但监管机构对 e 值的熟悉程度有限。沟通策略应着重说明 e 值与熟悉指标之间的对应关系：始终有效的 $p$值（$p_t = 1/\sup_{s \le t} E_s$）、置信序列以及校准后的效应估计。在申报材料中，可将 e 值监测作为一种预先计划的序贯分析方法来呈现，该方法采用单一的时间一致阈值，类似于组序贯监测，但不受固定查看时间表的限制。FDA医疗器械贝叶斯指南（美国食品药品监督管理局，2025年）为监管审评中的非传统证据框架提供了有益先例；2026年1月药品审评与研究中心/生物制品评价与研究中心关于药品和生物制品贝叶斯方法学的指南草案（美国食品药品监督管理局，2026年）则通过明确三种成功标准路径——I类错误校准、直接后验解释和决策理论方法——大幅扩展了监管框架，其中任何一种都可适用于 e 值监测。

有两个反复出现的误区值得强调。第一个是反复监测固定研究时限下的 $p$值，却未进行序贯校正，第二个是误以为后验概率阈值（例如，超过0.975）在可选停止下自然就是安全的；正如我们的数值研究所示（第 Section 10节），这两种做法都可能导致第一类错误大幅膨胀。第三个误区是 e 值特有的：不慎选择投注策略或混合备择；若 e 过程针对的治疗效应远大于或小于真实效应，检验效能可能会严重不足。良好实践要求明确记录滤过，包括每次中期观察时可用的信息，以及允许根据既往信息作出的适应性决策。e 过程的构造应被视为一项设计选择，须进行功效评估，通常通过模拟完成，这与组序贯设计中的边界选择遵循相同原则。

## 3 实用方案：实施 e值监测方案

实施工作首先要明确决策陈述，包括原假设 $H_0$、与估计目标对齐的终点以及分析人群。随后，通过确定 $\alpha$ 并决定采用单侧还是双侧监测来选择监测错误预算；期中决策将映射到如下形式的阈值 $E_t \geq 1/\alpha$，并在多臂情形下进行适当的逐族错误率或多重性调整。e过程的构建取决于终点。对于二元结局的双臂试验，第 Section5.4节所述的投注鞅提供了一种直接的构建方法（由 `eprocess_binary()` 在 `evalinger` 软件包中实现）。对于生存终点，安全对数秩检验提供了一种有原则的选择（Grünwald 等人 2021年）（由 `eprocess_logrank()`实现）。当具备部分先验信息但又希望保持稳健性时，正则化e过程在效率与稳定性之间提供了一种可调节的折中方案（Martin 2025年）。

报告方式应预先规定，包括数据安全监察委员会将审查 $E_t$、$\log E_t$ 还是推导得到的始终有效置信序列，以及哪些情况构成可采取行动的疗效、无效性或安全性信号。应在原假设和具有临床意义的备择假设下评估操作特征，以量化检验效能、期望样本量以及对设定错误的敏感性。

与组序贯设计不同，组序贯设计的效能计算基于信息分数和消耗函数，而基于 e 值的设计则需要考量 e过程在备择假设下的预期对数增长率。对于投注鞅 $E_n = \prod_{i=1}^{n}(1 + \lambda D_i)$，在设计备择假设 $(p_T, p_C)$ 下的期望停止时间可近似为 $\tau \approx \log(1/\alpha)/g(\lambda;\ p_T, p_C)$，其中 $g(\lambda;\ p_T, p_C) = \mathbb{E}_{H_1}[\log(1 + \lambda D_i)]$ 是每次观测的预期对数增长率。请注意，该增长率取决于个体应答率 $p_T$ 和 $p_C$，而不仅仅是二者之差 $\delta = p_T - p_C$：即使两个备择假设具有相同的 $\delta$，但基准率不同，其增长率也会不同，因为不一致结局的概率（只有这类患者配对才会产生证据）取决于基准率。对于固定的设计备择假设，采用增长率最优（GROW）的 $\lambda$ 选择可使增长率最大化。实际设计时，应针对一系列 $\lambda$ 值和临床上合理的备择假设，计算 $g(\lambda;\ p_T, p_C)$，选择能使设计备择假设下的检验效能最大化（或期望样本量最小化）的 $\lambda$，并通过模拟进行验证；`evalinger` 函数 `grow_lambda()`、`grow_lambda_grid()` 和 `edesign_binary()` 可自动完成这一流程。最大样本量 $N_{\max}$ 应设置为：在设计备择假设下，e过程有较高概率在 $N_{\max}$ 之前越过 $1/\alpha$，建议目标检验效能为 80–90%。与 `evalinger` 捆绑提供的交互式网络应用包含一个设计计算器，可将 $g(\lambda;\ p_T, p_C)$ 和期望停止时间可视化为下注比例的函数，从而无需编写自定义模拟代码即可快速探索设计空间。

## 4 背景：确证性试验中的成组序贯监测

反复查看固定时程检验统计量会增加假阳性的概率，除非该统计量和决策规则经过明确设计，能够保证其随时间推移仍然有效。这一现象促成了序贯分析的发展，其开端包括 Wald序贯概率比检验、Robbins 的序贯设计框架（Robbins 1952），以及它们在临床试验中的扩展（Wald，1947年；Armitage 1954, 1960）。成组序贯方法通过将信息累积划分为有限次期中分析，并构建界值以控制总体 I 类错误率，将这些理念付诸实践（Pocock，1977年；O’Brien 和 Fleming，1979年；Lan 和 DeMets，1983年；Jennison 和 Turnbull，2000年）。成组序贯检验通常假设，中期观察时的一系列检验统计量服从多元正态分布，其相关结构由信息分数决定。经典界值包括 Pocock 和 O’Brien–Fleming 界值。α 消耗函数（Lan 和 DeMets，1983年）通过根据信息时间分配累积第一类错误率，对这些界值进行了推广，从而允许灵活安排查看时点。

## 5 e值与e过程：简明教程

本节为具有临床试验背景的读者提供一份关于 e值和 e过程的独立入门介绍。有关全面论述，请参阅 Ramdas 等人（2023年）、Vovk 和 Wang（2021年），以及 Grünwald、Heide 和 Koolen（2023年）。

### 5.1 什么是 e 值?

一个 **e值 e** 是一个非负统计量 $E \geq 0$，其在原假设下的期望值至多为 1 ：

$$\mathbb{E}_{H_0}[E] \leq 1.$$

这就是完整的定义。“e”代表“证据”（其预期用途）或“期望”（其定义约束）（Vovk 和 Wang，2021年）。若要将 e值用作检验，则当 $H_0$ 满足 $E \geq 1/\alpha$时拒绝原假设。根据马尔可夫不等式，这样做可将第一类错误控制在水平 $\alpha$：

$$\mathbb{P}_{H_0}\left(E \geq \frac{1}{\alpha}\right) \leq \alpha.$$

对于水平为 $\alpha = 0.025$ 的单侧检验（这是确证性临床试验的标准），当 $E \geq 40$时我们拒绝原假设。e值为 50 比 e值为 25 提供了更强的证据，正如更小的 $p$值能提供更强的证据一样。与 $p$值不同，e值在许多情形下可以进行简单且有效的 组合运算：在独立（或更一般地说，给定既往信息时具有条件有效性）的情况下，e值的乘积仍是一个 e值；在任意依赖的情况下，算术平均仍是一个 e值（Vovk 和 Wang，2021年；Ramdas 等人，2023年）。

### 5.2 有限样本有效性

马尔可夫不等式提供的是精确的有限样本保证。该界 $\mathbb{P}_{H_0}(E \geq 1/\alpha) \leq \alpha$ 对任意样本量以及原假设下的任意分布均成立，无需渐近近似。在序贯情境下，e过程的相应时间一致保证由 Ville 不等式给出（第 Section 5.6节）：若$(E_t)$ 在 $H_0$下是非负上鞅，则 $\mathbb{P}_{H_0}(\sup_t E_t \geq 1/\alpha) \leq \alpha$。

对于临床试验，这一区别在三个情境中会带来实际影响。第一，**小样本和罕见病**：儿科试验、孤儿适应症以及细胞与基因治疗研究通常每组入组数十名患者，而非数百名。在这些样本量下，二项检验统计量的正态近似可能较差，而基于检验统计量渐近联合分布推导的组序贯界值可能无法维持其名义属性（Jennison 和 Turnbull，2000年）。e值保证无需修改即可成立。第二，**早期中期分析**：即使试验计划入组人数较多，首次中期观察也可能发生在患者人数相对较少时。在一项每组 200 人的试验中，信息分数为 0.1 时，观察发生于 $n = 20$——在这一阶段，渐近正态性值得怀疑，尤其是对于应答率接近 0 或 1 的二元终点。e 值监测无需依赖分布近似，即可在这一观察时点实现精确的第一类错误控制。第三，**具有复杂数据依赖关系的自适应设计**：反应自适应随机化、生物标志物驱动富集以及具有共享对照组的平台试验会产生相关结构，使标准序贯方法所依据的渐近论证变得复杂。只要 e 过程在原假设下仍为超鞅，e 值的有限样本保证便不受这些复杂性的影响。

当然，对于某些检验，$p$值也 可以在有限样本中实现精确有效——置换检验和精确二项检验就是典型例子。但要使 $p$值精确有效，就必须从范围受限的检验类别中进行选择，而且往往要牺牲检验效能或灵活性。根据定义，e值可“免费”获得有限样本有效性：在原假设下，任何期望值不超过 1 的非负统计量都立即是有效的 e值。在序贯情境下，这一区别更加明显：借助非负超鞅的 Ville不等式，e值能够在每个停止时刻同时提供随时有效的保证；而 $p$值通常要求样本量固定，或采用 α消耗等必须预先指定的校正方法（Lan 和 DeMets，1983年；Jennison 和 Turnbull，2000年）。

### 5.3 与你已熟悉的概念建立联系

临床试验生物统计学家熟练掌握成组序贯监测、α消耗、条件效能和贝叶斯后验概率。e值可自然地对应到上述每一项：

| 你熟悉的概念 | 对应的 e 值 |
|---|---|
| 中期分析时的检验统计量（例如，$Z$分数） | 患者 $n$时的 e值 $E_n$ |
| 拒绝边界（例如，O’Brien–Fleming $Z_k$） | 固定阈值 $1/\alpha$（例如，单侧 $\alpha = 0.025$ 为 40） |
| α 消耗函数 $\alpha^*(t)$ | 不需要——阈值在所有观察次数中保持恒定 |
| 预先指定的期中查看次数 | 不需要——可按需随时观察 |
| 设计备择假设（例如，$\delta = 0.15$） | 用于选择下注比例的设计备择假设 $\lambda$ |
| $H_1$下的期望样本量 | $N \approx \log(1/\alpha)/g(\lambda)$，其中 $g(\lambda)$ 是期望的每对患者的对数增长 |
| 中期分析时的条件效能 | 当前的 e值表示累积的证据；$g(\lambda)$ 表示未来证据增长的速度 |

| 你熟悉的概念 | e值对应概念 |
|---|---|
| 后验概率 Pr($\delta > 0$ \| 数据) | e值属于频率学派，但与贝叶斯因子有关（见下文） |
| 固定分析中的$p$值 | $1/E$ 始终是有效的 $p$值，具有任意时刻有效性<br>适用于序贯监测（见下文） |

关键的简化在于，**拒绝边界不会改变**，从一次观察到下一次都是如此。在组序贯设计中，你使用α 消耗函数，将第一类错误分配到预先计划的期中分析中：早期观察采用保守界值（较高的 $Z$分数），后期观察则更为宽松。使用e值时，无论何时观察、观察多少次，阈值始终保持在 $1/\alpha$。使这一切成为可能的数学机制——超鞅性质——会自动处理这种“消耗”。

### 5.4 运作原理：投注解读

构造临床试验 e值最直观的方法是采用**押注式检验**框架（Shafer，2021年；Shafer 和 Vovk，2019年）。想象一下，你以 1 美元起步，随着患者不断入组，反复押注反对原假设：

- **每对患者之后**，你观察治疗组或对照组哪一组效果更好，并拿出当前财富的 $\lambda$ 一部分下注，押治疗组更优。

- **如果原假设成立**（即不存在治疗效应），每次投注都是公平的：平均而言，你既不会赢钱，也不会输钱。
- **如果治疗组确实更优**，你的投注往往会有回报，财富也会随时间增长。
- **你当前的财富就是 e值**。当它达到 \$40（即 $1/\alpha$ for $\alpha = 0.025$）时，你积累的证据足以拒绝原假设。

为便于说明，我们将试验描述为产生患者配对（每组各有一个结局），这样每一步都能直接比较治疗组和对照组。这种配对并非必需：每当有新的结局信息到来时（包括各组入组人数不平衡时），都可以更新 e过程，前提是所构造的更新在原假设下始终保持为非负超鞅，且下一次更新所采用的投注选择是可预测的（即根据过去的信息可测）。

形式上，对于采用 1:1随机化且结局为二元的试验，定义 $D_i = X_i^T - X_i^C$为第 $i$ 个患者对。在 $H_0$：$p_T = p_C$ 的条件下，无论共同应答率是多少，我们都有 $\mathbb{E}[D_i] = 0$。投注者在经过 $n$ 个患者对（即每组 $n$ 个结局）后的财富为

$$E_n = \prod_{i=1}^{n} (1 + \lambda_i D_i),$$

其中，$\lambda_i \in (0,\ 1)$ 是在看到 $D_i$之前 选定的下注比例。端点不包括在内：$\lambda_i = 0$ 不会产生证据，而 $\lambda_i = 1$ 则会在 $D_i = -1$ 时永久令财富归零。由于 $D_i \in \{-1,\ 0,\ 1\}$ 且 $\lambda_i < 1$，每个因子 $(1 + \lambda_i D_i)$ 都严格为正。在 $H_0$下，我们有 $\mathbb{E}[D_i \mid D_1, \ldots, D_{i-1}] = 0$，因此 $(E_n)$ 是非负鞅（因而是 e过程；下文将给出精确定义）。对临床读者而言，有两点至关重要：

1. **原假设是复合的**——共同应答率 $p_T = p_C = p$ 是一个干扰参数。下注构造会自动处理这一点，因为对于 $p$的每一个 取值，$\mathbb{E}[D_i] = 0$，无需条件化、分层或估计。这类似于置换检验无需知道底层分布也有效，但此处的有效性还延伸到了序贯监测。

2. **下注比例 $\lambda$ 扮演着设计备择假设的角色**。设计组序贯试验的生物统计学家必须为效能计算指定一个预期效应量（例如，$\delta = p_T - p_C = 0.15$）。同样，选择 $\lambda$ 意味着你预期每对患者能提供多少证据。“合适的” $\lambda$ 取决于预期效应量，正如“合适的”样本量也取决于它（见下文“选择下注比例”）。

### 5.5 一个具体示例

假设我们为一种新的肿瘤治疗设计一项试验，预期应答率为 $p_T = 0.35$ （治疗组）对比 $p_C = 0.20$（对照组）。在采用 O’Brien–Fleming 界值并计划进行三次期中分析的标准组序贯设计中，我们通常会选择一个最大样本量，以便在 $\alpha = 0.025$（单侧）下达到目标检验效能（例如，80%）。

采用 e 值监测时，我们选择 GROW 最优下注比例 $\lambda^* \approx 0.37$ （推导见下文），并在每组患者对之后跟踪财富过程 $E_n$。考虑一种可能的轨迹（示例数字）：

- 50 组配对后： $E_{50} = 3.1$（低于 40；继续）
- 经过100对患者后： $E_{100} = 11.7$（持续增长，但低于40；继续）
- 经过120对患者后： $E_{120} = 45.2$（超过 $1/0.025 = 40$；拒绝 $H_0$）

试验可以在第120对患者时停止，也可以在 e 值超过阈值的任何其他时点停止。如果数据安全监察委员会还想查看第75对和第90对患者的数据（计划外），对于 e过程决策规则而言仍然有效，无需额外校正。在设计备择假设下，达到拒绝所需的患者配对数期望约为 $\log(40)/g(\lambda^*) \approx 131$对，同时仍保留不受限制的监测灵活性。

### 5.6 e值作为序贯检验统计量：e过程

一个**e过程**就是随着数据累积而计算的一系列 e 值（$E_1$， $E_2$，.....），其性质是：无论在任何时点停止——无论是计划内、计划外，还是由数据决定——都会得到一个有效的 e 值。下注构造所产生的财富过程 $E_n = \prod_{i=1}^{n}(1 + \lambda_i D_i)$ 是一个 e过程。其数学保证是 Ville不等式（Ville，1939年）：

$$\mathbb{P}_{H_0}(E_n \geq 1/\alpha \text{ at any time } n) \leq \alpha.$$

这相当于组序贯方法中的 α 消耗保证，但它更强：它对所有可能的停止规则都同时成立，而不只是对预先指定的一组中期观察成立。这种普适性的代价是：当查看时间表预先固定时，组序贯边界可能比单一的时间一致的 e 值阈值具有更高的检验效能；我们的数值研究量化了这一权衡。

对于熟悉贝叶斯监测的临床读者：e过程是一种频率学派工具，无需基于模拟进行校准即可控制第一类错误。贝叶斯后验概率阈值（例如，$\Pr(\delta > 0 \mid \text{data}) > 0.95$）在任意停止情形下不会自动控制第一类错误——必须针对每种特定设计，通过大量模拟进行校准（D. A. Berry，2025；美国食品药品监督管理局，2026）。e过程从构造上就提供了这一保证。

### 5.7 选择下注比例：设计备择假设的作用

在组序贯设计中，试验统计师会指定一个设计备择假设（即临床意义效应量），以计算达到目标检验效能所需的样本量。e 值监测中的对应步骤是选择<用户下注比例 $\lambda$，以确定下注时反对原假设的激进程度。

在备择假设 $p_T > p_C$ 下，对于 $\lambda \in (0,1)$，每个患者配对贡献的期望对数证据为

$$g(\lambda) = p_T(1-p_C)\log(1+\lambda) + (1-p_T)p_C\log(1-\lambda),$$

其中，剩下的两种结局——双方均有响应（概率为 $p_T p_C$）和双方均无响应（概率为 $(1-p_T)(1-p_C)$）——产生 $D_i = 0$，因此对期望增长率的贡献为 $\log(1+\lambda\cdot 0) = 0$。这就是**增长率**：在备择假设下，每个患者配对（每组一次观测）平均获得的证据量（以对数单位计）。增长率最优（GROW）下注比例 $\lambda^*$ 使 $g(\lambda)$ 最大：

$$\lambda^* = \frac{p_T(1-p_C) - (1-p_T)p_C}{p_T(1-p_C) + (1-p_T)p_C}.$$

对于我们的示例（$p_T = 0.35, p_C = 0.20$），得到 $a = 0.35 \times 0.80 = 0.28, b = 0.65 \times 0.20 = 0.13$，因此 $\lambda^* = 0.15/0.41 \approx 0.37$，增长率为 $g(\lambda^*) \approx 0.028$ 纳特/患者配对。预期达到拒绝所需的患者配对数为

$$N_{\text{expected}} \approx \frac{\log(1/\alpha)}{g(\lambda^*)} = \frac{\log(40)}{0.028} \approx 131,$$

这是效能计算中期望样本量对应的 e 值类比量。关系式 $N \approx \log(1/\alpha)/g(\lambda)$是基本设计方程：效应量越大，增长率越高，因此试验所需时间越短，这与经典设计完全一致。

如果真实效应大于预期，e值会增长得更快，试验也会更早停止。如果真实效应较小，e值增长缓慢——这时下注比例就很重要。选择过大的 $\lambda$（过度下注），就像针对一个并不存在的大效应为试验设定检验效能：e值会来回波动，可能永远无法越过阈值。选择 $\lambda$

过小的 $\lambda$（投注不足），就像开展一项检验效能过高的试验：最终会拒绝原假设，但会浪费患者。在实践中，可以先根据设计备择假设设定 $\lambda$，再利用不断累积的数据对其进行自适应调整，前提是适应规则具有可预测性（即仅使用观察下一个结果之前可获得的信息）。

### 5.8 p 值与贝叶斯因子的关系

任何 e 值的倒数都是有效的 p 值，符合通常的固定时限意义：对于任何预先指定的$\alpha \in (0,\ 1)$，$\mathbb{P}_{H_0}(p' \leq \alpha) = \mathbb{P}_{H_0}(E \geq 1/\alpha) \leq \alpha$。在序贯监测中，通过 $p_t = 1/\sup_{s \leq t} E_s$可得到始终有效的 p 值过程，该过程在可选停止时控制第一类错误（Vovk 和 Wang，2021年；Shafer，2021年）。然而，与所有 p 值一样，在看到数据之后选择拒绝水平 $\alpha$ 通常并非有效的推断操作；在实践中，最好将 e 值视为主要证据尺度，并在需要时将其转换为 p 值，作为报告层。校准函数也可将标准 p 值转换为 e 值，但所得 e 值的检验效能通常低于专门构造的 e 值（Vovk 和 Wang，2021年）。

E 值也与**贝叶斯因子**有关。当原假设是简单的（即只有一个分布）时，任何贝叶斯因子都是有效的 e值。当原假设是复合假设时——大多数临床试验都属于这种情况——两者的关系就更为微妙：大多数贝叶斯因子并非e值，而大多数 e值也不是贝叶斯因子（Ramdas 等，2023年；Polson、Sokolov 和 Zantedeschi 2026年；Datta 等人 2025年）。混合方法构造（用于生存终点的安全对数秩检验；Grünwald 等人（2021年））所产生的 e值与贝叶斯因子相似——它们会对备择假设上的某个混合分布进行似然积分——但选择该混合分布的目的是最大化证据的增长，而非反映先验信念。对于贝叶斯试验学家而言，这是一种熟悉的数学形式，同时还具有频率学派有效性保证。

### 5.9 构建 E 值的三种方法

除了下注构造（我们将其用于二元终点）之外，在临床环境中还有另外两种实用方法（Ramdas 等，2023年；Grünwald、Heide 和 Koolen 2023年）：

**混合方法**。指定一个针对合理治疗效应的“混合”分布 $W$，并根据该分布对似然比求平均：$E = \int L(\theta)\ dW(\theta)$，其中 $L(\theta)$ 是针对特定效应量 $\theta$ 的似然比。这类似于计算贝叶斯因子，但选择混合分布是为了提高统计效率，而非体现先验信念。针对至事件时间终点的**安全对数秩检验**采用了这种方法，其混合分布以预期风险比为中心（Grünwald 等人 2021年）。对于熟悉贝叶斯分析的试验统计学家而言，这种构造是最熟悉的构建 e值的方法。

**校准器（p值到e值的转换）**。如果你已经通过标准分析得到一个有效的 $p$值 $P$，则可以应用一个校准器函数 $f$：$[0,1] \to [0,\infty)$，该函数单调不增且满足 $\int_0^1 f(u)\ du \leq 1$，并令 $E = f(P)$，从而得到一个 e值。这样便能以简单的方式为现有检验添加一层 e值，但与基于似然或下注构造得出的专门构造的 e值相比，效率会有所降低（Vovk 和 Wang 2021年；Ramdas 等，2023年）。

贝叶斯因子是两个模型下边际似然的似然比；若以简单原假设为基准构造，或满足 e值期望上界，则可作为 e值。然而，自适应试验中使用的贝叶斯决策规则通常依赖后验概率或预测概率是否超过固定阈值（例如，$\mathbb{P}(p_T > p_C \mid \text{数据}) > 0.95$）。如果不进行校准，这类阈值在可选停止情况下并不能自动实现频率学派错误控制；校准通常通过大量模拟来完成（D. A. Berry，2025年；美国食品药品监督管理局，2025年、2026年）。相比之下，e值通过 Ville不等式直接给出解析形式的第一类错误保证，但要获得有竞争力的检验效能，可能需要仔细调优。

## 6 自适应临床试验的 e 值

### 6.1 期中监测、方案适应性调整和反应自适应随机化

在许多自适应试验中，期中分析的次数和时间并非严格固定（例如，由于信息驱动的数据安全监察委员会排期、延迟结局或操作限制）。只要 e过程在包含适应性决策的滤过下保持有效，E过程监测便允许任意增加观察次数，而无需重新推导消耗边界。无论进行两次还是二十次期中观察，阈值 $E_t \geq 1/\alpha$ 都不会改变，也无需指定或重新校准 α 消耗函数。这使得在单一阈值下对安全性或疗效进行连续监测成为可能。

样本量修改、试验组剔除、富集或反应自适应随机化（RAR）等适应通常都是既往信息的函数。用鞅的术语来说，这些属于可预测的设计选择。只要相对于扩大的滤过，e过程仍为超鞅，这类可预测适应就不会影响 e过程的有效性。

反应自适应随机化（RAR）对经典方法和基于 e 值的方法都构成了特殊挑战。Berry 将 RAR 的思想源流追溯至 Thompson 1933 年的提议（Thompson 1933；D. A. Berry，2025）：该提议按照当前某组优于其他组的贝叶斯概率，将下一位患者分配至该组。用 Berry 的话说，这一策略“比 Hill 的发表早了 15 年”，而且“比 Hill 的固定随机分配更灵活，也可以说更有用”。Thompson 抽样已在现代平台试验中大规模应用：在 I-SPY 2 中，自适应随机化使更多患者在其所属分子亚型内被分配至前景看好的试验组。Berry 报告称，奈拉替尼的病理完全缓解率比对照组高出 17 个百分点，其中 10 个百分点的提升归因于自适应随机化，而非药物本身的作用（Barker 等人，2009；Park 等人，2016；D. A. Berry，2025）。对于 RAR 下的 e过程，关键要求是下注比例 $\lambda_i$必须可预测——即在观察 $D_i$之前确定——但它可以依赖于用于患者 $i$的分配概率，而这些概率本身是既往数据的函数。在最简单的情形下（固定 $\lambda$），这一要求自动得到满足。对于自适应 $\lambda_i$，则需谨慎确保 $\lambda_i$ 的适应规则和 RAR 规则相对于同一滤过都是可预测的。Thompson 抽样与下注检验传统具有共同的结构：二者都通过计算依赖当前后验分布的量来作出决策（分配治疗或累积证据），都具有内在的序贯性，也都因其构造方式而满足可预测性。近期关于反应自适应设计精确频率学派分析的研究（Baas、Jacko 和 Villar，2025）补充了 e 值方法，提供了计算上可行的精确检验，可用作校准基准。

### 6.2 至事件时间终点与安全对数秩检验

对于生存终点，序贯监测已是标准做法，而安全对数秩检验则提供了一种基于 e 值的对应方法，在可选停止和继续的情况下仍能保证第一类错误率（Grünwald 等人，2021年）。其构造过程如下：在每个观测到的事件时间，对数秩贡献是治疗组中观察到的事件数与根据风险集计算的预期事件数之差。在风险率相等的原假设下，该贡献的均值为零，这为构造下注型 e 过程提供了依据。安全对数秩检验通过对对数风险比的先验进行积分，构造混合似然比，从而得到一个可随事件累积而逐步计算的 e 值。 `safestats` R 软件包（Schure和Ly，2022年）提供了相应的实施工具，可用于指定设计备择假设（以风险比和先验集中度表示）、计算期望停止时间，以及为达到目标检验效能校准先验。

与标准组序贯对数秩检验相比，安全对数秩检验在任何固定分析时点都会牺牲一些检验效能，但它可以在超过计划的最大时限后继续收集数据，而不会使推断失效；当事件入组速度慢于预期，或外部情况导致计划中的分析延期时，这一特性在实际操作中很有价值。生存试验中一个实际问题是结局延迟：当主要终点为总生存期时，在每次期中分析时，许多患者虽已接受治疗，却尚未发生事件。Berry 在 ASTIN、AWARD-5、GBM AGILE 和 Leqembi 201 阿尔茨海默病试验中采用了应对这一问题的方法：建立纵向模型，根据早期测量预测晚期终点，并使用贝叶斯多重插补，将部分随访数据纳入期中决策（Geiger 等，2012；Alexander 等，2018；D. A. Berry 等，2023；D. A. Berry，2025）。对于 e 过程方法，结局延迟会造成滤过间隙：在时间 $t$ 的 e 过程理想情况下应纳入时间 $t$可获得的所有信息，但部分结局数据存在非随机缺失。将 e 值与贝叶斯纵向建模相结合的一种途径是，在每个事件时间利用当前最佳的结局预测来构造下注增量；只要插补模型具有可预测性，原假设下的 e 过程就仍然有效。

### 6.3 通过 e过程进行无效性监测

无效性停止是疗效停止的补充，出于伦理和经济原因同样重要。e过程通过两种途径支持无效性判断。

第一种途径使用置信序列。在每次中期观察时，对 e过程进行反演，以获得治疗效应的$(1-\alpha)$ 置信序列 $C_n$ $\delta = p_T - p_C$ ( 见第 7 节 )。如果 $C_n$ 的上界低于最小临床重要差异 $\delta_{\min}$，则可以停止试验，判定其无效；此时的证据水平足以表明，数据与具有临床意义的效应不相容。由于置信序列在任意时点均有效，该无效性规则无需预先设定无效性评估的次数或时间。

第二种途径　　构建一个独立的 e过程，用于检验“反转 的”原假设

$$H_0' : p_T - p_C \geq \delta_{\min}.$$

定义中心化增量

$$D_i' = \delta_{\min} - (X_i^T - X_i^C).$$

在 $H_0'$下，我们有

$$\mathbb{E}[D_i'] = \delta_{\min} - (p_T - p_C) \leq 0,$$

因此， $D_i'$ 的均值非正。具有 $\lambda' \in (0,\, 1/(1-\delta_{\min})]$ （此上界确保对所有 $D_i' \in [\delta_{\min}-1,\, \delta_{\min}+1]$， $1+\lambda' D_i' \geq 0$ ）的投注鞅 $E_n' = \prod_{i=1}^{n}(1+\lambda' D_i')$在 $H_0'$下是非负上鞅。请注意， $D_i'$ 仅取决于 $\delta_{\min}$ （预先指定的设计参数）和观测数据，而不取决于任何未知的干扰参数。当真实效应小于 $\delta_{\min}$时， $\mathbb{E}[D_i'] > 0$ 且财富过程增长，不断累积支持无效性的证据。如果 $E_n'$ 超过 $1/\alpha_f$ （其中 $\alpha_f$ 是无效性错误预算），则有强有力的证据表明治疗组未能达到具有临床意义的效应。

### 6.4 多重性与平台试验

平台试验会监测多个试验组、亚群和终点中的多个假设。与 $p$值类似方法相比，e值支持代数形式更简单的组合运算：对于同一原假设，e值的算术平均在任意依赖关系下仍是有效的 e值（Vovk 和 Wang 2021；Ramdas 等，2023年），而独立 e值的乘积仍是有效的 e值。这为通过专为 e值设计的程序（例如 e-BH）控制错误发现率提供了支持；在 Wang 和 Ramdas（2022）所述条件下，这类程序具有正式的保证。
具体而言，考虑一项采用共享对照组和 $K$ 个试验组的平台试验——I-SPY 2、GBM AGILE、REMAP-CAP/COVID 和 Precision Promise 均采用了这种架构（Barker 等人，2009；Park 等人，2016；Alexander 等，2018；Angus 等人，2020；D. A. Berry，2025）。Woodcock 和 LaVange（Woodcock 和 LaVange，2017）对这类主方案的分类进行了形式化，并以 I-SPY 2 作为典型示例。针对每个试验组 $k$，都会维护一个投注 e过程 $E_n^{(k)}$，用于检验 $H_0^{(k)}$： $p_k = p_C$。当满足 $E_n^{(k)} \geq 1/\alpha_k$ 时，试验组晋级；达到无效性标准时，则将其淘汰。试验组被淘汰后，其 e过程将被冻结；新试验组加入时，则会初始化一个新的 e过程。由于每个 e过程都针对自身假设提供任意时点有效的 level-$\alpha_k$ 检验，因此，通过并集界可以将所有晋级试验组的家族错误率控制在 $\sum_k \alpha_k$ （其中 $\alpha_k$ 分配给各试验组）。共享对照会导致假设之间存在依赖关系；如果设计目标是在多个试验组或亚组中控制错误发现率，则可使用基于 E 值的错误发现率程序（例如 e-BH）筛选 e值，并在 Wang 和 Ramdas（2022）所述条件下获得保证。 `evalinger` 函数 `platform_monitor()` 和 `ebh()` 可在此情境下实现多试验组 E 过程监测及 e-BH 多重性控制。平台试验还可能通过层级建模和时间趋势调整，借用非同期对照组或其他来源的信息（Saville 等人，2022；D. A. Berry，2025）。这种借用会改变监测问题的依赖结构；因此，要保证 e过程有效，借用机制及任何自适应加权都必须相对于监测滤过是可预测的。FDA 贝叶斯指南草案讨论了平台试验情境下的非同期对照组和借用考量，包括有效样本量核算，以及针对先验与数据冲突的敏感性分析（美国食品药品监督管理局，2026；Morita、Thall 和 Müller，2008；Evans 和 Moshonov，2006）。在此类情境中，e过程可作为额外的、对时间一致的证据核查，与用于设计和论证借用模型的贝叶斯运行特性分析并行使用。

## 7 报告、估计与校准

临床试验报告通常围绕一些熟悉的对象展开： $p$值、置信区间和预先设定的决策阈值。当使用 e过程进行监测时，将 $p_t = 1/\sup_{s \leq t} E_s$设为 $\mathbb{P}_{H_0}(\inf_t p_t \leq \alpha) \leq \alpha$，即可得到始终有效的 $p$值过程；根据 Ville不等式，该过程满足相应性质（Vovk 和 Wang 2021；Shafer，2021年）。当试验的沟通场景需要报告 $p$值时，这些转换很有用；但应将其视为底层 e过程之上的报告层，而非主要的设计对象。
序贯监测很少只涉及检验；它还涉及在每次中期观察时估计治疗效应，并量化不确定性。通过检验反演，e过程可以得到始终有效置信序列。对于用于表示治疗效应的参数 $\theta$，针对每个假设值构造一个 e过程 $E_t(\theta)$，并定义$C_t = \{\theta : \sup_{s \leq t} E_s(\theta) < 1/\alpha\}$。在满足正则条件时， $C_t$ 是一个 $(1-\alpha)$ 置信序列：真实参数曾经落在 $C_t$ 之外的概率至多为 $\alpha$ （Howard 等人 2021；Ramdas 等，2023年）。这类序列提供了置信区间在时间一致意义上的类比，并可与监测决策一同报告，以便在期中分析中解读效应量估计。

e值的实际优势并不意味着不再需要设计校准，而是改变了校准的目标。第一类错误保证是解析性的，不依赖于消耗计划，但检验效能和期望样本量取决于 e过程的构造方式。在参数设置中，这种依赖体现为对备择模型或投注策略的选择；在非参数设置中，则可能体现为正则化方案的选择。因此，设计任务从选择 α消耗边界转变为选择并调优 e过程，并通过在临床上合理的情景中进行模拟来评估其表现（Grünwald、Heide 和 Koolen 2023年；Martin 2025年）。

## 8 可选延续与证据综合

e值与经典检验输出之间的一个重要区别在于，e值可以在一系列自适应确定的实验中进行有意义的累积。如果每个组成部分都针对同一个零假设主张产生有效的 e值 $E^{(k)}$，那么在满足标准的独立性或条件独立性条件时，即使 $K$ 取决于数据，且研究 $k+1$ 的设计是在观察到先前结局后才确定的，乘积 $E^{(1)}E^{(2)}\cdots E^{(K)}$ 仍然是一个 e 值（Grünwald、Heide 和 Koolen 2023年；Ramdas 等人 2023年）。这一性质为序贯荟萃分析思维提供了形式化途径：随着时间推移、跨越不同研究阶段积累数据，可以不断更新证据，而不会破坏第一类错误控制。

在临床开发中，乘积累积应被视为一种有原则的证据框架，而非自动合并分析的许可，因为当人群、终点或操作性定义发生变化时，各项研究中的零假设主张很少会完全相同。实际应用需要仔细阐明所检验的主张，包括如何在各组成部分之间定义估计目标，以及允许何种形式的变异。尽管如此，在固定时限分析容易导致非正式且易出错的证据汇总的情境下，可选延续这一性质提供了一种数学上连贯的方法，可用于随时间追踪证据。

## 9 实施、软件与监管背景

由于 e 过程保证具有时间一致性，操作性监测计划可以简化：中期查看可以在任意时间进行，但监测统计量和阈值保持不变。监测章程应明确说明滤过，包括每次期中分析纳入哪些数据流和协变量、如何处理延迟结局，以及允许基于既往信息的可预测函数作出哪些适应性决策。计算应与决策权限分开：e过程根据不断累积的数据进行计算，但停止、继续或修改试验的决定仍属于治理行动，必须在临床和伦理层面得到论证。在每次中期观察时报告 $\log E_t$，可形成一条加法型证据轨迹，比一系列二元边界穿越更易于解读；与置信序列结合时，还能同时传达证据强度以及效应量的不确定性。

用于临床试验中 e值构造的软件日趋成熟。 `safestats` R软件包（Schure和Ly 2022年）为两组比较（比例、均值和生存）提供安全检验，并配备设计校准工具（期望停止时间、所需样本量和检验效能），包括安全对数秩检验、安全 $z$检验和安全 $t$检验，还可以针对指定的设计备择假设计算 GROW 最优下注比例。Ramdas 等人（2023年）配套提供的研究代码实现了额外的 e 过程构造。临床试验模拟平台（例如， `rpact`、`gsDesign`）尚未原生支持 e值监测，但可以将 e 过程计算嵌入自定义模拟框架中。

为弥合 e 值文献与临床试验实际运行之间的差距，我们提供了 `evalingerR`软件包（源代码和文档见 https://github.com/VadimSokolov/evalinger；可通过 `devtools::install_github("VadimSokolov/evalinger")`安装），该软件包实现了本文所述的完整方法。该软件包提供：（i） `eprocess_binary()`和 `eprocess_logrank()`，用于构造二元和生存终点的 e过程；（ii） `grow_lambda()`、 `expected_growth_rate()`和`edesign_binary()`，用于GROW最优设计校准；（iii） `emonitor()`，配合流式 `update()`，用于实时序贯监测；（iv） `confseq_binary()`，用于时间一致的置信序列；（v） `futility_cs()`和 `futility_eprocess()`，通过置信序列和倒数 e过程两种途径进行无效性监测；（vi） `platform_monitor()` 和 `ebh()`，用于采用 e-Benjamini–Hochberg 多重性控制的多臂平台试验监测；（vii） `hybrid_monitor()`，用于联合 e过程和组序贯监测；以及（viii）`simulate_comparison()`，用于数值演示部分所报告的五种方法的蒙特卡洛比较。该软件包可与成熟的 R 临床试验生态系统（`gsDesign`、 `rpact`、 `survival`、`safestats`）集成。软件包附带并托管于 https://sailtargets.shinyapp s.io/evalinger/ 的交互式 Shiny 网络应用提供三个仪表板——设计计算器、监测仪表板和方法比较——使临床试验人员无需编写代码即可探索该方法：试验人员可以针对其设计备择假设校准 GROW最优下注比例，按批次模拟监测并实时可视化 e过程轨迹和置信序列，还可以在用户指定的情景下，对 e值、组序贯和贝叶斯监测规则进行直接比较。该网络应用支持美国食品药品监督管理局 2026年FDA贝叶斯指导草案对贝叶斯和适应性设计所要求的运行特征文档编制（美国食品药品监督管理局 2026年）。

从监管角度来看，美国食品药品监督管理局关于药品和生物制品自适应设计的指南（美国食品药品监督管理局，2019年）要求自适应方法以强意义上的方式控制第一类错误，并维护试验完整性。e值在构造上满足错误控制要求（Ville不等式），但该指南并未专门提及e值，监管审评人员也可能不熟悉这种方法。Berry 三十多年来将贝叶斯方法引入监管实践的经验，是一个相关的先例（D. A. Berry，2025）。这条

道路始于医疗器械贝叶斯上市前审批（包括约在2000–2001年的一项脊柱植入物试验，以及 Pennello 和 Thompson（2007）与 Irony 和 Campbell（2011）记录的后续器械与放射健康中心审评经验），历经器械与放射健康中心的器械贝叶斯指南 [（最初于2010年发布；美国食品药品监督管理局（2025）；Irony 等（2023））]，直至药品审评与研究中心接受完全采用贝叶斯方法的注册试验（2003年的 Pravigard Pac、2014年的 Trulicity/AWARD-5（Geiger 等，2012）、Leqembi 阿尔茨海默病项目（D. A. Berry 等，2023），以及作为无缝分期注册平台的 GBM AGILE（Alexander 等，2018））。这一过程既不平坦，也不迅速。Berry 指出，这需要监管倡导者、大量基于模拟的运行特征评估，以及同时以贝叶斯和频率学派术语报告结果。Berry 报告称，在 JAMA 编辑因 ThermoCool AF 试验未提供频率学派报告而提出异议后，他现在会在任何被要求提供的频率学派指标旁注明：“本文中的所有 $p$值和置信区间均为描述性结果，不具有推断意义”（D. A. Berry，2025）。

这一发展轨迹中的一项里程碑式进展，是2026年1月由药品审评与研究中心和生物制品评价与研究中心联合发布的、关于在药物和生物制品临床试验中使用贝叶斯方法学的FDA 指南草案（美国食品药品监督管理局2026年）。该指南将数十年的监管经验正式确立；正如Viele在Berry Consultants的评论中所指出的，它代表着“贝叶斯临床试验与监管科学的一次巨大飞跃”（Viele 2026年）。该指南有三项特征与我们在此发展的方法学直接相关。首先，该指南规范了贝叶斯试验中成功标准的三条不同路径：（i）针对I 类错误率进行校准，即通过模拟调校贝叶斯后验阈值，以达到单侧2.5%的家族错误率——这是迄今美国食品药品监督管理局大多数贝叶斯试验所采用的框架，也是我们的校准的贝叶斯规则在第10节中所示的方法；（ii）对后验概率作直接解释，即各方同意先验“能够准确概括当前的信念状态”，并由后验概率本身为决策提供依据，而无需进行频率学派校准；（iii）获益–风险或决策理论方法，其中成功阈值纳入明确的损失函数，以权衡批准无效药物与未能批准有效药物的后果。e值最自然地对应第一条路径，它们为I 类错误控制提供解析保证，从而避免指南对后验阈值方法所要求的大量基于模拟的校准。但它们也补充了第二和第三条路径：与纯粹的贝叶斯决策规则并行运行的e过程，可提供一份频率学派证据账本，从而使尚未准备接受直接后验解释的监管机构或审评人员感到放心。

其次，指南草案将分析先验（用于最终分析）与设计先验（用于在一系列合理情景下评估操作特征）区分开来（Spiegelhalter、Abrams 和 Myles，2004年；美国食品药品监督管理局，2026年）。这与 e过程的设计校准相类似：选择投注策略（或等效的调优参数），使其在设计备择假设下具有较高效率，然后在一系列备择假设下评估敏感性（如表3所示）。指南草案强调，在小样本以及包含早期期中分析的设计中，操作特征可能对先验假设尤为敏感；在这些情形下，密集或不规则监测会使基于模拟的校准更加脆弱。

第三，指南草案讨论了借用以及先验与数据冲突，强调通过模拟在设计阶段探索操作特征和敏感性分析（Evans 和 Moshonov，2006年；美国食品药品监督管理局，2026年）。这一背景与借用情形下的 e值相关，因为 e值可以在以所借用数据为条件（将其视为滤过的一部分）的情况下构造，并能够提供任意时刻有效的 I 类错误控制，而无需在监测安排发生变化时重新推导序贯阈值。在使用非同期对照借用的平台试验中——包括指南草案所提及的情形——若某个 e过程在其滤过中考虑了借用机制，则可提供时间一致的证据保证，从而补充监管机构所要求的贝叶斯运行特征说明（美国食品药品监督管理局，2026年）。

e值支持者面临着与 Berry 三十多年前所应对的类似转化挑战。实用的沟通策略包括：将 e 值监测方案表述为一种预先计划的序贯分析方法，并设定单一的时间一致阈值；在 e 过程之外，同时提供始终有效的 $p$ 值和置信序列，以便采用熟悉的方式进行报告；并在统计分析计划中纳入基于模拟的运行特征（检验效能、期望样本量、第一类错误），这与组序贯设计和贝叶斯设计的惯例一致。美国食品药品监督管理局的复杂创新设计（CID）倡议鼓励采用新型试验设计来证明有效性（美国食品药品监督管理局，2019年）；而2026年贝叶斯指南草案对非传统成功标准持开放态度，这或可为确证性试验中的基于 e 值的监测提供一条监管途径。

## 10 数值演示

本节比较双臂随机对照试验中针对二分类终点的五种监测规则。其目的是说明错误控制和检验效能如何取决于推断对象与停止规则，并

量化不同监测方法之间的操作权衡。所有结果均由配套脚本`run_simulations.R` 和 `run_extended_sims.R` 使用 `evalinger` 包生成，因此完全可复现。[1]

### 10.1 设计与监测规则

我们考虑一项双臂随机对照试验，采用 1:1随机化，将患者分配至治疗组 (T) 和对照组 (C)。每位患者的结局均为独立的伯努利抽样： $X_i^T \sim \text{Ber}(p_T)$ 和 $X_i^C \sim \text{Ber}(p_C)$。我们检验

$$H_0 : p_T = p_C \quad \text{versus} \quad H_1 : p_T > p_C,$$

在两种假设下均为 $p_C = 0.3$，在 $H_0$ 下为 $p_T = 0.3$，并且在 $H_1$ 下为 $p_T = 0.45$（治疗效应为 15 个百分点）。每组的最大样本量为 $N_{\max} = 200$（总计 400 人）。我们进行 20 次等间隔的期中分析，每组每增加 10 名患者进行一次。单侧第一类错误的名义目标为 $\alpha = 0.025$。

该设计备择假设下的 GROW 最优下注比例为 $\lambda^* = 0.312$，期望增长率为$g(\lambda^*) = 0.0238$纳特/患者配对；在备择假设下，预计约 155 对时停止。

五种监测规则如下：

1. **朴素重复 $p$-值**：每次中期观察时，针对 $p_T > p_C$ 计算单侧 Wald $z$检验；当 $p$值低于 $\alpha$时，试验停止。
2. **校准的组序贯方法**：将同一 $z$统计量与 O'Brien–Fleming 型边界 $z \geq c/\sqrt{t}$进行比较，其中 $t = n/N_{\max}$ 为信息分数， $c$ 通过蒙特卡洛校准，以达到总体第一类错误率 $\alpha$。3. **E值投注法**：使用 GROW 最优 $\lambda$的投注鞅$E_n = \prod_{i=1}^{n}(1 + \lambda D_i)$。当$E_n \geq 1/\alpha$时拒绝。4. **朴素后验阈值**：采用 Jeffreys 先验计算贝叶斯后验概率 $\Pr(p_T > p_C \mid \text{数据})$，当其超过 $1 - \alpha = 0.975$时停止。5. **校准的贝叶斯方法**：使用相同的后验分布，但通过模拟校准阈值，以将第一类错误率控制在 $\alpha$ （S. M. Berry 等，2010年；美国食品药品监督管理局，2025年、2026年）。

所有五种规则均已在 `evalinger`函数 `simulate_comparison()` 中实现。

### 10.2 结果

表2 报告了基于 50,000 次模拟试验，在以下条件下得到的蒙特卡洛估计： $H_0$ 和 $H_1$。对于每种规则，我们报告拒绝概率（在 $H_0$ 下为第一类错误，在 $H_1$ 下为检验效能），以及每组的平均停止样本量。

表2：在二组二分类随机对照试验中比较五种监测规则（$p_C = 0.30$，设计备择假设 $p_T = 0.45$，每组 $N_{\max} = 200$，20 次观察， $\alpha = 0.025$）。拒绝概率和平均样本量（每组）根据 50,000 次蒙特卡罗重复模拟估计。

| 监测规则 | 原假设下的拒绝次数 | 备择假设下的拒绝次数 | 原假设下的平均样本量 | 备择假设下的平均样本量 |
|---|---|---|---|---|
| E值投注法（$\lambda = 0.31$） | 0.012 | 0.723 | 199.013 | 139.230 |
| 校准的组序贯方法（OBF） | 0.025 | 0.861 | 198.930 | 139.819 |
| 朴素的重复 $p$ 值 | 0.148 | 0.933 | 178.376 | 74.206 |
| 朴素后验阈值 (0.975) | 0.135 | 0.932 | 181.013 | 76.233 |
| 校准的贝叶斯方法 | 0.020 | 0.688 | 197.085 | 133.913 |

朴素的重复 $p$ 值规则和朴素后验概率规则在频繁进行中期分析时，都会出现明显的第一类错误膨胀。校准的组序贯边界通过设计控制第一类错误，并且在错误控制方法中检验效能最高。校准的贝叶斯规则也能控制第一类错误（通过模拟校准阈值构造而成），其显著优势是在备择假设下可以更早停止。本次模拟中，E 值规则以较大余量控制第一类错误，反映出这一特定 e过程的保守性（Ville不等式不一定是紧的）。

图 1 展示了 对数 $E_n$ 在原假设 $H_0$ 和 $H_1$下 12 次模拟试验的样本路径。原假设成立时，对数 E 过程呈下降趋势（原假设为真时，投注者平均会损失财富）。备择假设下，样本路径趋于上升，大多数在 $N_{\max}$之前就越过阈值，展现了反对错误原假设的证据如何不断累积。

[1]要重新生成结果： `Rscript run_simulations.R && Rscript run_extended_sims.R` 从论文目录执行。输出保存到 `sims/`。

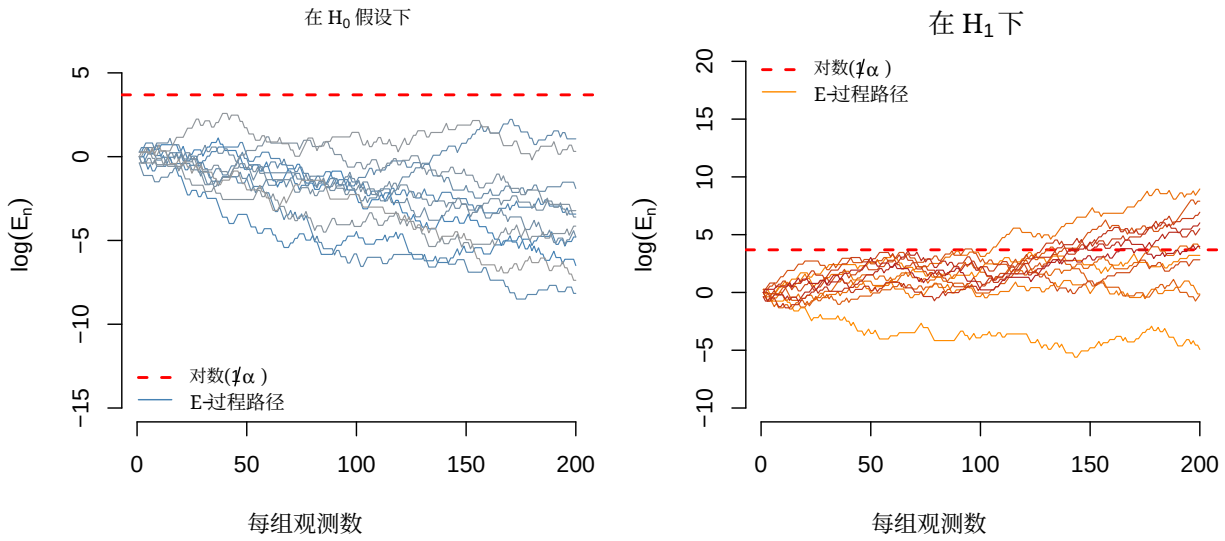


图 1：12 次原假设（左）和备择假设（右）下模拟试验的对数 E 过程 $\log E_n$ 的样本路径。水平虚线标示拒绝阈值 $\log(1/\alpha)$。在 $H_0$ 下，投注者的对数财富逐渐下降；在 $H_1$ 下，大多数路径在达到最大样本量之前越过阈值。

### 10.3 对下注比例的敏感性

e 值监测规则取决于下注比例 $\lambda$，其作用类似于组序贯设计中消耗函数的选择，或贝叶斯监测中的先验。表 3 展示了检验效能和期望样本量如何随 $\lambda$ 变化。

表3：e 值监测对下注比例 $\lambda$ 的敏感性。对于所有选择，第一类错误率均低于 $\alpha = 0.025$，但检验效能差异显著。GROW 最优 $\lambda$ 可使检验效能最大化，而保守或激进的选择则会显著降低检验效能。

| $\lambda$ | 原假设下的拒绝次数 | 备择假设下的拒绝次数 | 原假设下的平均样本量 | 备择假设下的平均样本量 |
|---|---|---|---|---|
| 0.10 | 0.000 | 0.119 | 200.000 | 197.495 |
| 0.20 | 0.004 | 0.642 | 199.828 | 161.593 |
| 0.31 | 0.012 | 0.722 | 199.048 | 139.825 |
| 0.40 | 0.014 | 0.685 | 198.521 | 135.428 |
| 0.50 | 0.014 | 0.589 | 198.190 | 139.088 |

保守的选择（$\lambda = 0.10$）几乎没有检验效能，因为单位观测增长率太慢，无法在 200 次观测内达到拒绝阈值。GROW 最优的 $\lambda \approx 0.31$ 可使检验效能最大化。激进的选择（$\lambda \geq 0.40$）会牺牲检验效能，因为对数 E 过程的方差增加，其影响超过了期望漂移的提高。在所有情况下，第一类错误率都远低于 $\alpha$，这证实了任意时刻有效控制对下注比例具有稳健性。这项敏感性分析表明，将 $\lambda$ 校准至设计备择假设至关重要，这与组序贯设计校准消耗函数、贝叶斯设计校准先验的理念一致。

理论增长率图景（图2）直观展示了为何 GROW 最优 $\lambda$ 至关重要：过度下注（$\lambda$ 过大）可能会使增长率变为负值，而投注不足则会浪费证据。

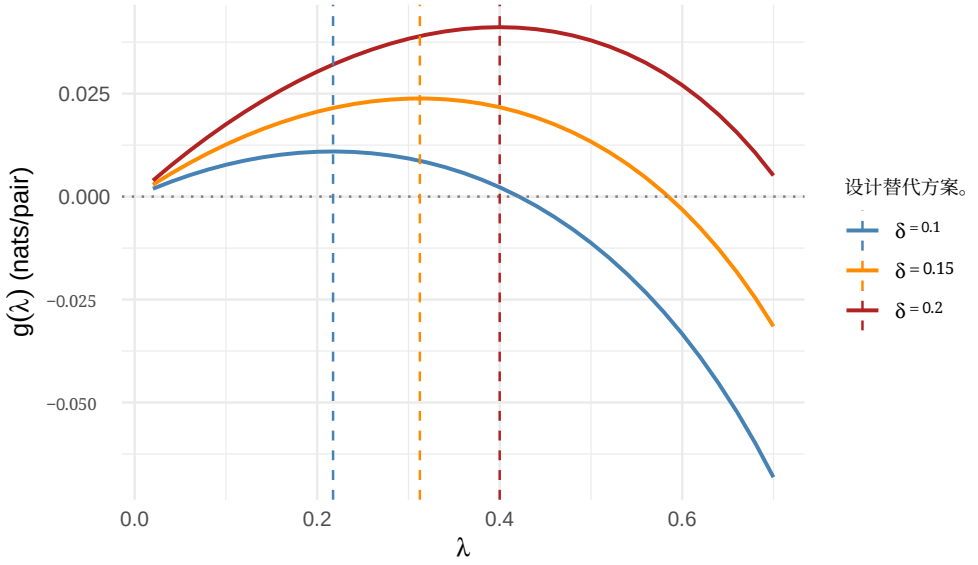


图2：预期对数增长率 $g(\lambda;\ p_T,\ p_C)$ 随下注比例 $\lambda$ 变化的情况，针对三种设计备择假设（$\delta = p_T - p_C = 0.10$，0.15，0.20，且 $p_C = 0.30$）。垂直虚线标示每种备择假设对应的 GROW 最优 $\lambda^*$。

### 10.4 混合监测

在实践中，试验可以考虑使用相同的数据，同时运行组序贯边界和 e过程，将组序贯规则作为主要决策标准，并将 e过程作为补充性证据摘要。表4 展示了在一个 GROW 最优 $H_1$模拟试验中，每次中期观察时评估的两条监测流。

表4：$H_1$下单次模拟试验的混合监测表。在每次中期观察时，都会评估 e过程和组序贯边界。始终有效的 $p$值为 $1/\max_{s\le n} E_s$。

| Look | n | 信息分数 | $\hat{\delta}$ | $z$ | GS 界 | GS 拒绝 | 日志 $E$ | E 拒绝 | AV $p$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 0.05 | 0.400 | 1.952 | 9.594 | 假 | 0.985 | 假 | 0.373 |
| 2 | 20 | 0.10 | 0.350 | 2.394 | 6.784 | 假 | 1.801 | 假 | 0.165 |
| 3 | 30 | 0.15 | 0.333 | 2.802 | 5.539 | 假 | 2.514 | 假 | 0.081 |
| 4 | 40 | 0.20 | 0.325 | 3.128 | 4.797 | 假 | 3.124 | 假 | 0.036 |
| 5 | 50 | 0.25 | 0.360 | 3.951 | 4.290 | 假 | 4.484 | TRUE | 0.011 |
| 6 | 60 | 0.30 | 0.333 | 4.038 | 3.917 | TRUE | 5.028 | TRUE | 0.007 |
| 7 | 70 | 0.35 | 0.314 | 4.007 | 3.626 | TRUE | 5.366 | TRUE | 0.003 |
| 8 | 80 | 0.40 | 0.262 | 3.576 | 3.392 | TRUE | 4.889 | TRUE | 0.003 |
| 9 | 90 | 0.45 | 0.267 | 3.891 | 3.198 | TRUE | 5.704 | TRUE | 0.003 |
| 10 | 100 | 0.50 | 0.270 | 4.131 | 3.034 | TRUE | 6.315 | TRUE | 0.002 |
| 11 | 110 | 0.55 | 0.273 | 4.384 | 2.893 | TRUE | 7.130 | TRUE | 0.001 |
| 12 | 120 | 0.60 | 0.258 | 4.299 | 2.769 | TRUE | 7.300 | TRUE | 0.000 |
| 13 | 130 | 0.65 | 0.254 | 4.367 | 2.661 | TRUE | 7.638 | TRUE | 0.000 |
| 14 | 140 | 0.70 | 0.286 | 5.118 | 2.564 | TRUE | 9.541 | TRUE | 0.000 |
| 15 | 150 | 0.75 | 0.267 | 4.914 | 2.477 | TRUE | 9.130 | TRUE | 0.000 |
| 16 | 160 | 0.80 | 0.269 | 5.125 | 2.398 | TRUE | 9.843 | TRUE | 0.000 |
| 17 | 170 | 0.85 | 0.259 | 5.094 | 2.327 | TRUE | 9.910 | TRUE | 0.000 |
| 18 | 180 | 0.90 | 0.250 | 5.040 | 2.261 | TRUE | 9.874 | TRUE | 0.000 |
| 19 | 190 | 0.95 | 0.258 | 5.352 | 2.201 | TRUE | 10.961 | TRUE | 0.000 |
| 20 | 200 | 1.00 | 0.240 | 5.089 | 2.145 | TRUE | 10.278 | TRUE | 0.000 |

图 3 展示了共享时间轴上的 e过程和组序贯边界。两种监测流提供了互补信息：组序贯边界在离散查看时给出二元信号，而 e过程则提供连续的证据轨迹。

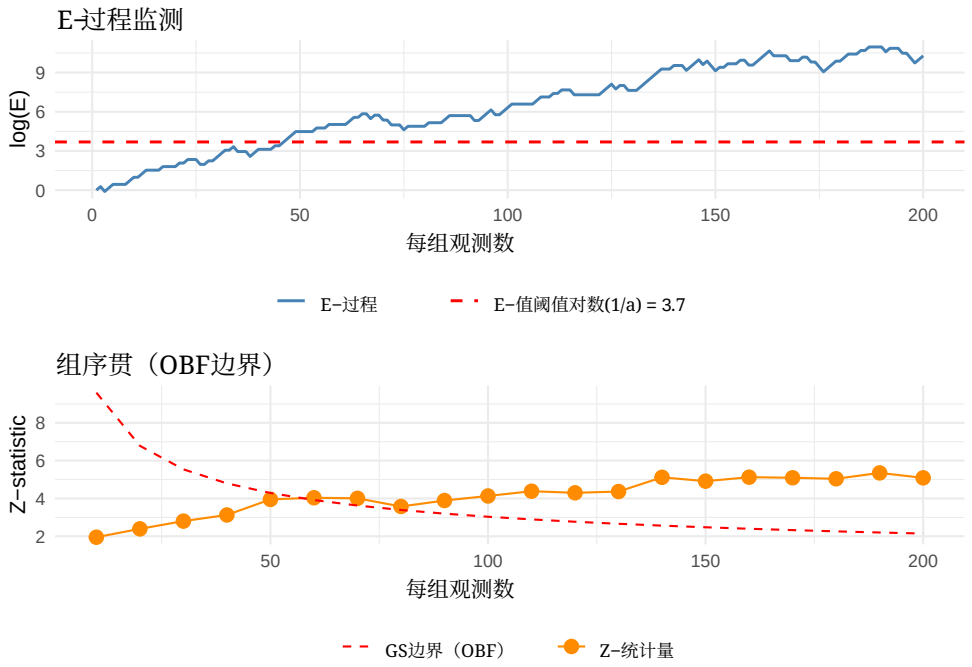


图3：在 $H_1$下对单项试验进行的混合监测。上图：对数 E 过程及其拒绝阈值。下图：每次查看时的 $z$统计量与O'Brien–Fleming边界。两条轨迹均越过各自的阈值，但时机与轨迹有所不同。

在 $H_1$下进行的 50,000 次模拟试验中，组序贯规则与 E 值规则在 72.0% 的试验中结果一致（两者均拒绝），在 13.7% 的试验中也结果一致（两者均不拒绝）。在 14.1% 的试验中，组序贯规则拒绝而 e值不拒绝；仅在 0.2% 的试验中，e值拒绝而组序贯规则不拒绝。这种一致性模式证实，e 值监测通常比经过校准的组序贯监测更保守，但它能提供有用的证据摘要，且无论监测计划如何都有效。

### 10.5 监测计划对功效差距的影响

E值方法的一项关键主张是，其保证不受监测计划影响。为在实际操作中检验这一点，我们比较了四种察看时点安排下的 e值监测与重新校准的组序贯监测：5 个固定观察时点、20 个固定观察时点、5 个随机时间观察时点（对 200 个随机时间表取平均），以及连续监测（每对患者都进行观察）。对于每种安排，我们都通过蒙特卡洛重新校准 OBF边界常数 $c$，使组序贯规则恰好达到 $\alpha = 0.025$ 的第一类错误率；e值阈值 $1/\alpha = 40$ 则始终保持不变。

表5：四种监测安排下e值与重新校准的组序贯检验效能的比较。无论监测安排如何，e值阈值均固定为 $1/\alpha$。GS边界针对每种安排重新校准。在连续监测下，GS边界变得极为保守（检验效能为10%），而e值检验效能增至75%。

| 时间表 | 方法 | 第一类错误 | 检验效能 | 备择假设下的平均样本量 |
|---|---|---|---|---|
| 固定（5 次观察） | e值 | 0.008 | 0.686 | 154.140 |
| 固定（5次观察） | GS（重新校准） | 0.025 | 0.870 | 149.738 |
| 固定（20次观察） | e值 | 0.012 | 0.723 | 139.230 |
| 固定（20次观察） | GS（重新校准） | 0.025 | 0.861 | 139.819 |
| 不规则（5 次随机观察） | e值 | 0.007 | 0.679 | 160.365 |
| 不规则（5 次随机观察） | GS（重新校准） | 0.025 | 0.868 | 157.669 |
| 连续（200 次观察） | e值 | 0.016 | 0.750 | 131.056 |
| 连续（200 次观察） | GS（重新校准） | 0.043 | 0.100 | 180.192 |

表 5 中的结果呈现出清晰的规律。在观察次数较少的固定日程下，组序贯规则明显优于 e值（5 次观察时，检验效能为 87%，而 e值为 69%）。观察次数增加到 20 次后，GS 的优势略有缩小（86% 对 72%）。然而，在连续监测——也就是 e值旨在发挥优势的情形——下，即使经过重新校准，GS边界的检验效能也只有 10.0%，而 e值的检验效能为 75.0%。这是因为，为了在大量观察时点下维持第一类错误率，OBF边界 $z \geq c/\sqrt{t}$ 必须提高，最终在早期观察时点变得过于保守，以至于实际上无法触及。相比之下，e值阈值保持不变。在不规则时序（随机查看时间）下，e值的检验效能保持稳定，而 GS 规则则需要针对每个实际生成的时间表重新校准——这是 e值完全避免的一项操作复杂性。图4 直观展示了这些结果。

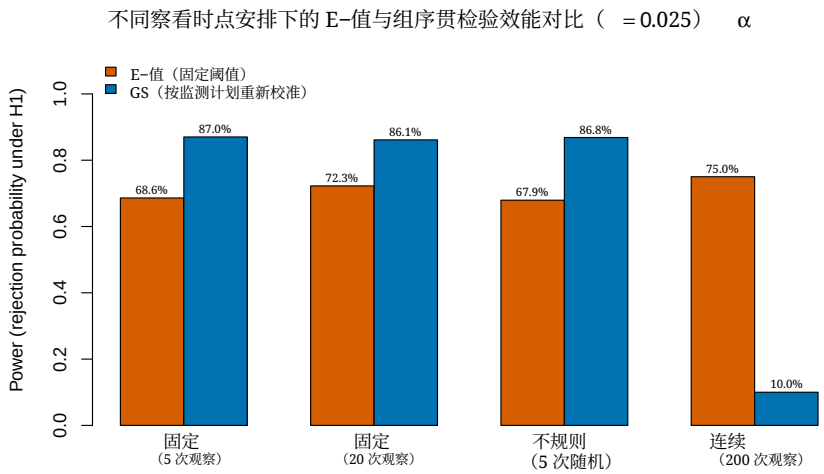


图4：四种监测安排下的 e 值与组序贯检验效能。e值阈值固定；针对每种监测安排重新校准 GS 边界。在连续监测下，GS 检验效能骤降，而 e值检验效能上升。

### 10.6 不同设计设置下的参数敏感性

e值与组序贯方法之间的功效差距并非恒定：它取决于基线率 $p_C$、效应量 $\delta$ 和监测强度。表6报告了18种设计配置网格下的 e 值和 GS 检验效能（$p_C \in \{0.10, 0.30, 0.50\}$、$\delta \in \{0.10, 0.15, 0.20\}$、$K \in \{5, 20\}$ 次观察），每种配置均基于20,000次蒙特卡罗重复模拟。

表6：e值与 不同设计配置下的组序贯检验效能（每组 $N_{\max} = 200$，$\alpha = 0.025$）。该功效差距最大的是 在中等基线率下的小效应中最小，在低基线率下的大效应中最大。

| $p_C$ | $\delta$ | 观察次数 | $\lambda^*$ | e值 | GS | Gap |
|---|---|---|---|---|---|---|
| 0.1 | 0.10 | 5 | 0.385 | 0.583 | 0.795 | 0.211 |
| 0.3 | 0.10 | 5 | 0.217 | 0.262 | 0.547 | 0.285 |
| 0.5 | 0.10 | 5 | 0.200 | 0.223 | 0.499 | 0.276 |
| 0.1 | 0.15 | 5 | 0.500 | 0.891 | 0.977 | 0.086 |
| 0.3 | 0.15 | 5 | 0.312 | 0.685 | 0.864 | 0.179 |
| 0.5 | 0.15 | 5 | 0.300 | 0.659 | 0.840 | 0.181 |
| 0.1 | 0.20 | 5 | 0.588 | 0.978 | 0.999 | 0.021 |
| 0.3 | 0.20 | 5 | 0.400 | 0.900 | 0.980 | 0.080 |
| 0.5 | 0.20 | 5 | 0.400 | 0.900 | 0.978 | 0.077 |
| 0.1 | 0.10 | 20 | 0.385 | 0.619 | 0.785 | 0.166 |
| 0.3 | 0.10 | 20 | 0.217 | 0.292 | 0.533 | 0.241 |
| 0.5 | 0.10 | 20 | 0.200 | 0.250 | 0.494 | 0.244 |
| 0.1 | 0.15 | 20 | 0.500 | 0.908 | 0.974 | 0.066 |
| 0.3 | 0.15 | 20 | 0.312 | 0.718 | 0.856 | 0.139 |
| 0.5 | 0.15 | 20 | 0.300 | 0.696 | 0.844 | 0.149 |

表6：不同设计配置下的 e值和组序贯检验效能（每组 $N_{\max} = 200$， $\alpha = 0.025$）。在中等基线率下，小效应的功效差距最大；在低基线率下，大效应的功效差距最小。

| $p_C$ | $\delta$ | 观察次数 | $\lambda^*$ | e值 | GS | Gap |
|---|---|---|---|---|---|---|
| 0.1 | 0.20 | 20 | 0.588 | 0.983 | 0.999 | 0.016 |
| 0.3 | 0.20 | 20 | 0.400 | 0.918 | 0.979 | 0.060 |
| 0.5 | 0.20 | 20 | 0.400 | 0.917 | 0.979 | 0.062 |

出现了两种模式。首先，随着效应量增加，功效差距显著缩小：对于 $\delta = 0.20$ 有$p_C = 0.10$，差距仅为2.1个百分点（e值为97.8%，而GS为99.9%）；而对于 $\delta = 0.10$ 有 $p_C = 0.30$，差距则扩大到28.5个百分点。其次，将观察次数从5次增加到20次会适度缩小差距（通常缩小3–5个百分点），因为更频繁的监测使e值能够更早检测到阈值穿越，而GS边界必须将其α预算分配得更加分散。基线率 $p_C$ 也很重要：较低的基线率（每个患者对中不一致的配对更多）会使两种方法都获得更快的增长率和更高的检验效能，其中e值的相对获益更大。

### 10.7 无效性监测演示

为演示第 5.6 节所述的无效性监测方法，我们模拟一种情形：治疗组产生的效应虽小，但在临床上无关紧要：$p_T = 0.33$， $p_C = 0.3$ (真实 $\delta = 0.03$)，最小临床重要差异为 $\delta_{\min} = 0.1$。最大样本量为每组 300 人。

我们应用两种无效性判定路径：(i) 置信序列方法：当 95% CS 的上界低于 $\delta_{\min}$ 时判定为无效；(ii) 倒数 e 过程，在 $\alpha_f = 0.10$ 时检验 $H_0'$: $\delta \geq \delta_{\min}$。在 10,000 次模拟中，基于 CS 的方法在 14.2% 的试验中检出无效性（在检出的试验中，配对数中位数为 147.5），而倒数 e 过程在 53.5% 的试验中检出无效性（中位数为 108 对）。倒数 e 过程更敏感，因为它直接检验无效性假设；而置信序列路径要求整个置信带都排除 $\delta_{\min}$，这是一个更保守的判据。这两种机制均已在 `evalinger(futility_cs()` 和 `futility_eprocess())` 中实现。

图 5 展示了在单次试验中应用这两种方法的效果。左侧面板显示置信序列逐渐收窄并围绕真实效应（$\delta = 0.03$），直到其上界降至 $\delta_{\min} = 0.10$以下，从而触发无效性判定。右侧面板显示
随着证据不断累积，倒数 e 过程逐渐增大，表明证据不支持临床意义效应。

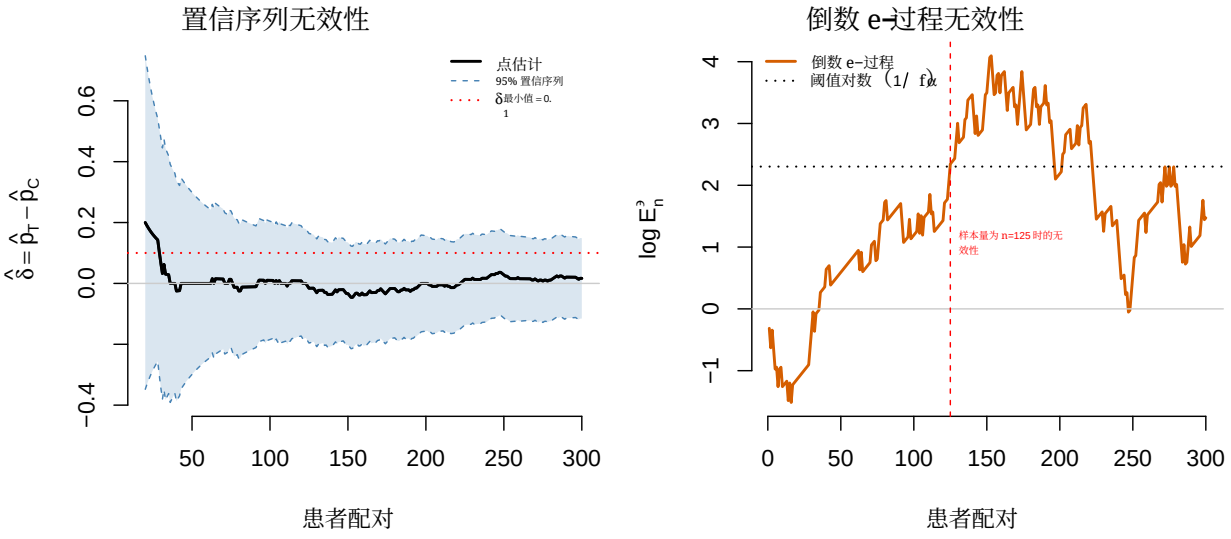


图 5：针对具有亚临床效应的试验进行无效性监测（$p_T = 0.33$， $p_C = 0.30$， $\delta_{\min} = 0.10$）。左：95% 置信序列与最小临床重要差异阈值；当上界低于 $\delta_{\min}$ 时，判定为无效。右：检验 $H_0'$: $\delta \geq \delta_{\min}$ 的倒数 e 过程；当该过程越过 $\log(1/\alpha_f)$ 时，判定为无效。

### 10.8 类 RECOVERY 大型试验模拟

为展示该方法在现代确证性试验常见规模下的应用，我们模拟了一项参照 RECOVERY 地塞米松组设计的试验（RECOVERY 协作组，2021年）。已发表结果显示，28天死亡率为 22.9%（地塞米松组）和 25.7%（常规护理组），绝对降低 2.8 个百分点。我们根据这些比率，模拟一项每组有 $N = 2000$ 名受试者的 1:1 随机试验。

GROW 最优下注比例为 $\lambda^* = 0.0760$，增长率为 $g(\lambda^*) = 0.001065$ 纳特/对；其速度比主要研究慢一个数量级（$g \approx 0.024$），反映出效应量小得多。期望停止时间约为 3462 对，超过了每组 2000 人的试验规模。在一次模拟试验中，e过程达到$E = 8.8$（低于阈值 $1/\alpha = 40$），死亡率差异的始终有效的95%置信序列为 $[-0.019, 0.075]$。在 10,000 次模拟中，该 e值在每组 $N = 2000$ 人时达到 31.4% 的检验效能；在拒绝原假设的试验中，拒绝的中位时间为 1417 对。

这项分析说明了两个重要观点。首先，对于较小的效应量（$\delta \approx 0.03$），要达到足够的检验效能，e值需要非常大的样本量——3462 对样本的期望停止时间远超典型试验能够招募的样本量。这正是群序贯方法优势最大的情形，因为它能够针对固定时间表优化界值。其次，即使 e 过程没有越过拒绝阈值，置信序列仍能提供有用信息：区间 $[-0.019, 0.075]$ 在任意时刻均有效，并且正确包含真实效应，可对证据作出易于解读的总结，供数据安全监察委员会结合群序贯决策使用。图6 展示了 e 过程轨迹和置信序列。

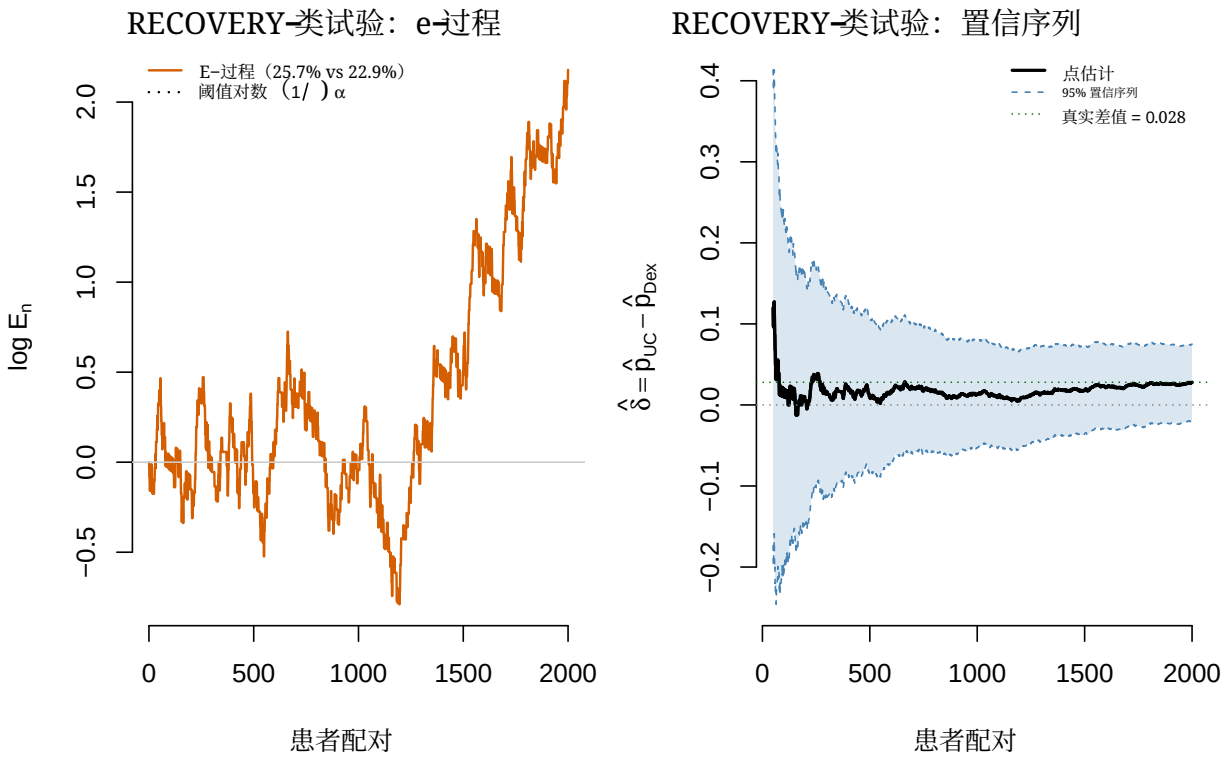


图 6：RECOVERY样试验模拟（$p_{UC} = 25.7\%$、$p_{Dex} = 22.9\%$、每组 $N = 2000$）。左：e 过程轨迹稳步上升，但在 2000 对样本内未达到拒绝阈值，这反映了效应量较小。右：95% 置信序列围绕真实效应收窄，并提供任意时刻有效估计。

## 11 真实数据示例：Novick（1965年）溃疡试验

为在真实临床试验数据上阐释这一方法，我们重新审视最早期的随机对照试验贝叶斯分析之一：Novick 和 Grizzle 于1965年开展的一项研究，比较了四种十二指肠溃疡手术治疗方案（Novick 和 Grizzle 1965年）。该试验在19家退伍军人管理局医院开展，将400名患者随机分配至治疗方案A、B、C和D（每组100人）。六个月时的结局分为优良、一般/较差或死亡。观察到的死亡人数为：A = 7，B = 1，C = 1，D = 3。

这组数据非常适合我们的目的，原因有三。第一，原论文采用两种方法分析数据：经典的受限序贯设计和贝叶斯后验可信度分析；加入e值视角后，便可进行自然的三方比较。第二，多臂结构使我们能够通过 `platform_monitor()` 和 `ebh()`展示平台监测和多重性控制。第三，Novick本人指出，对于序贯再评估，贝叶斯方法提供了“经典方法无法给予的一定程度的灵活性”——而e值正是通过时间一致保证将这类灵活性形式化。

### 11.1 E过程监测

我们将死亡结局视为二分类终点，并考察每种治疗的死亡率是否高于表现最佳的治疗 B。对于每一项两两比较，我们使用`eprocess_binary()` 构建投注鞅 e过程，并采用共同的 GROW 最优下注比例 $\lambda^* = 0.727$；该比例针对死亡率为 6% 对 1% 的设计备择假设进行校准（依据试验中观察到的死亡率；Novick 先验预期的 2–3% 基线死亡率与这些数值大体相符）。显著性水平为 $\alpha = 0.025$ （单侧）。
由于无法获得原始的患者水平序贯顺序，我们在每个组内模拟随机到达顺序（使用固定随机种子以确保结果可复现），并随着观察到每一对患者而构建 e过程。图 7展示了由此得到的证据轨迹。

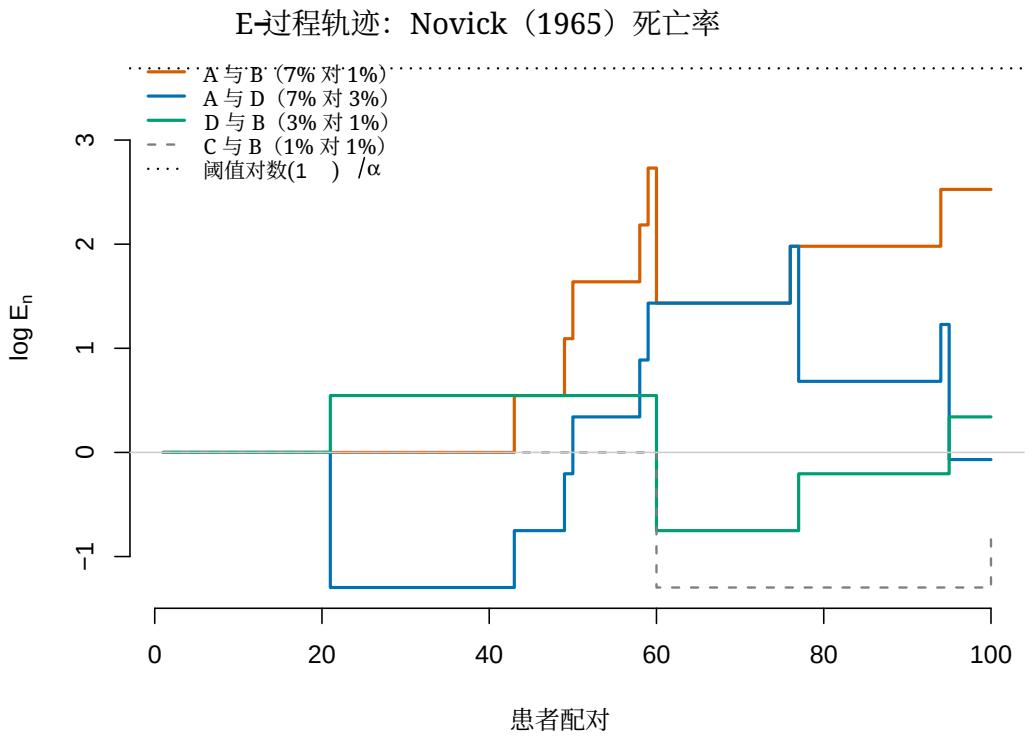


图7：Novick（1965）试验中死亡率的四组两两比较的e过程轨迹。水平虚线表示拒绝阈值 $\log(1/\alpha) = 3.69$。最强比较（A 对 B，7% 对 1%）累积了大量证据（$E \approx 15.3$），但未跨越阈值，这表明在每组有100名患者且事件罕见的情况下，任意时刻有效推断具有保守性。C 对 B 的比较（两组均为1%）则正确地显示没有证据累积。

表7 比较了 e 值证据与 Novick 最初报告的贝叶斯后验可信度。Novick的贝叶斯分析报告了可信度 $C(p_A > p_B) \approx 0.98$，用于比较 A 组与 B 组，显示死亡率差异的证据极强。相比之下，e过程得出 $E \approx 15.3$，对应的任意时刻有效 $p$值为 0.065——这是中等证据，未达到 $\alpha = 0.025$ 阈值。这种差异体现了任意时刻有效性的代价：无论数据被检查多少次，e值保证都成立；而贝叶斯可信度并不会自动控制重复监测下曾经宣布假阳性的概率。采用此处的设计备择假设时，期望停止时间约为 182 个患者配对，超过可用的 100 对，这证实该试验的统计功效不足，无法在此 $\alpha$ 水平上得出确凿的 e 值拒绝结论。

表7：Novick（1965）试验中死亡率的两两比较。 $E$ 是序贯监测期内达到的最大e值；AV $p$ 是始终有效的 $p$值（$1/$最大值 $E$）；Novick 可信度是在泊松模型和均匀先验下的后验概率 $C(\lambda_i/\lambda_j > 1)$（摘自 Novick 的表4）。

| 比较 | 死亡人数（组 $i$） | 死亡人数（组 $j$） | $E$ | AV $p$ | Novick 可信度 |
|---|---|---|---|---|---|
| A组与B组 | 7 | 1 | 15.3 | 0.065 | 0.980 |
| A 与 C | 7 | 1 | 45.8 | 0.022 | 0.980 |
| A 与 D | 7 | 3 | 7.2 | 0.138 | 0.887 |
| D 与 B | 3 | 1 | 1.7 | 0.579 | 0.813 |
| D 与 C | 3 | 1 | 5.2 | 0.194 | 0.813 |

### 11.2 置信序列与平台监测

治疗效应的任意时刻有效置信序列 $\delta = p_A - p_B$ 在 $n = 100$ 处为 $[-0.034, 0.154]$，点ˆ估计值为 $\delta = 0.06$（图 8）。由于此区间包含零，e 值分析得出的结论是一致的：数据并未提供任意时刻有效的证据来排除死亡率相等的可能性，尽管点估计显示两者相差 6 个百分点。

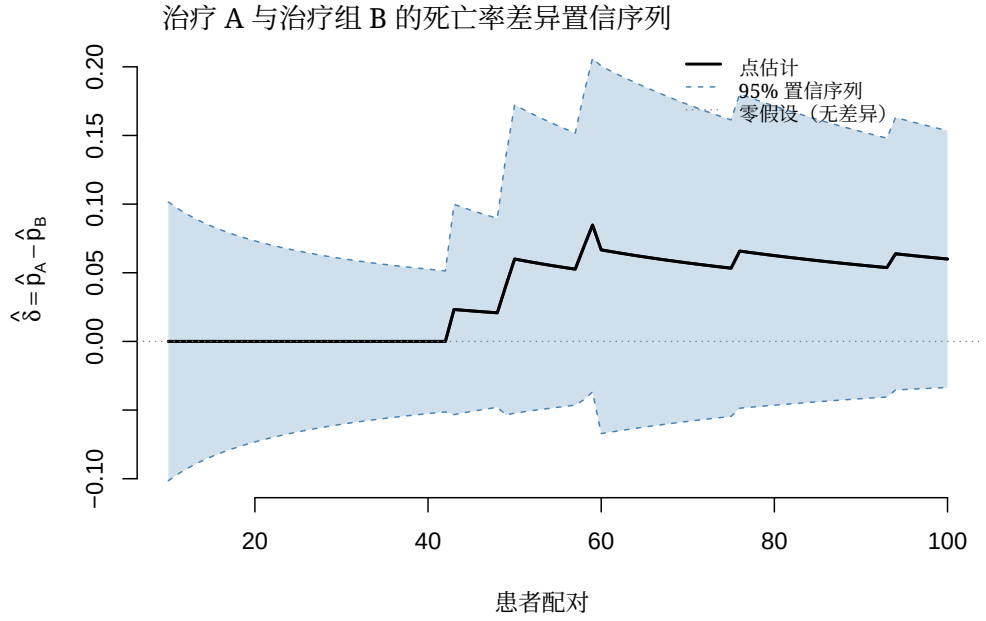


图 8：Novick（1965）试验中死亡率差异 $\delta = p_A - p_B$ 的始终有效的95%置信序列。随着数据不断累积，置信带逐渐缩小，但在 $n = 100$时仍包含零，这与 e过程未被拒绝的结果一致。

当将试验作为平台进行分析时（治疗 B 作为共享对照组，A、C、D 组作为试验组），在错误发现率水平为 0.05 的情况下，e-BH 程序在四次中期观察中的任何一次均未拒绝任何假设（$n = 25$，每组 50、75、100 名患者）。这体现了多重性控制所带来的额外保守性，并确认在事件发生率为 1–7% 时，每组 100 名患者不足以实现多重性调整后的拒绝——鉴于数据有限，这是一个恰当的结论。

跨分析范式的三方比较颇具启发性。Novick 的经典序贯分析（Armitage 的限制性序贯方案）在观察了 400 名患者后也未能达到决策边界：基于偏好的检验在六项比较中均未发现显著的两两差异，而 Novick 估计，在 A 与 B 的比较可能达到边界之前，还需要大约 200 个可绘制配对。e 值分析与这一频率学派结论一致：$E \approx 15.3$ 是有力证据，但不足以拒绝$\alpha = 0.025$。相比之下，Novick 的贝叶斯可信度为 0.98，是个例外——它暗示着近乎确定性，但在可选停止的情况下，这种高可信度并不具备频率学派的错误控制保证。这与本文提出的理论框架一致：e值是一种频率学派工具，在“何谓‘足够的证据’”这一点上与经典方法一致，

同时在灵活监测方面具有操作优势。与只能在预先计划的边界穿越时宣布“无显著差异”的经典序贯检验不同，e 值分析能够生成连续的证据轨迹（$E = 15.3$，对应始终有效的 $p$值 0.065）、随样本量增加而缩小且在每个样本量下均有效的置信序列，并保证无论研究人员检查累积数据多少次，这些汇总结果始终有效——而这正是 Novick 认为贝叶斯方法独有的灵活性。本分析的代码可在 `run_novick_analysis.R` 中获取，并使用 `evalinger`函数 `eprocess_binary()`、 `confseq_binary()`、 `platform_monitor()`和 `ebh()`。

## 12 讨论

e值提供了时间一致的证据接口，适合临床试验的序贯特性。与α消耗和组序贯监测相比，e过程监测在中期分析不规则或临机进行时简化了管理，并为可选延续提供了一种连贯机制。与贝叶斯后验或预测概率监测相比，e值无需大量基于模拟的校准，即可在可选停止下提供直接的频率学派保证，同时仍允许采用受贝叶斯启发的构造（混合分布、先验）来提高效率。

数值研究显示，在20次等间隔查看下，校准的组序贯边界（86.1%）与投注e过程（72.3%）之间的功效差距约为14个百分点。然而，扩展模拟（第10.5节–第10.8节）表明，这一差距高度依赖具体情境。在连续监测下，组序贯检验效能骤降至10%，而e值检验效能升至75%，因为为了在多次查看时控制第一类错误，OBF边界必须变得极为保守。参数网格（表6）显示，差距从2个百分点（大效应）到29个百分点（中等基线率下的小效应）不等；类RECOVERY模拟（第10.8节）证实，对于小效应（$\delta \approx 0.03$），e值所需样本量远超典型的试验入组规模。在固定日程下，有几种策略可以缩小这一差距：纳入累积信息的自适应投注比例 $\lambda_i$ 、正则化e过程（Martin 2025年），以及在预先计划的期中分析时使用组序贯边界、在计划外查看时使用e过程的混合方法。

校准的贝叶斯监测规则经过校准，可控制第一类错误（这是通过构造实现的），但施加校准后，其检验效能并不一定高于 e 值设计。在我们的示例中，校准的贝叶斯规则将第一类错误率控制在 2.0%，检验效能为 68.8%；在备择假设下，每组平均在 134 名受试者时停止。其缺点在于，校准依赖于特定模拟：约 0.998 的后验阈值是针对特定原假设（$p_T = p_C = 0.30$）、查看时间表（观察 20 次）和分析先验（每组采用杰弗里斯 Beta$(0.5,0.5)$ ）计算得出的；其中任何一项发生变化，都需要重新进行模拟。这种脆弱性是后验阈值监测在实践中的代价，与 e 值阈值 $1/\alpha$形成对比：一旦指定了有效的 e 过程，该阈值在可选停止下仍然有效，无需重新校准。2026年FDA贝叶斯指导草案（美国食品药品监督管理局，2026年）从监管视角阐明了这一权衡，区分了分析先验与设计先验：校准阈值取决于分析先验，而表2中的功效评估则对应于设计备择假设下的操作特征。该指南要求在“各种合理的设计先验”下评估操作特征，这意味着必须针对一系列原假设和备择假设情景重新审视校准；随着试验复杂度增加，模拟负担也会加重。相比之下，对于任何停止规则，e值都能通过解析方法提供第一类错误保证；设计任务主要转为通过选择下注比例来优化检验效能，并进行敏感性分析（如表3所示）。

E值并非经典工具的通用替代品。其实际表现取决于所选的投注策略，而设计不当的 e过程可能会严重缺乏检验效能。对于复杂终点——例如具有非比例风险的生存时间结局、复发事件，或存在信息性缺失的纵向结局——如何构建兼具高检验效能与稳健性的 e过程，仍是一个前沿研究课题。Berry 的经验说明了这一点：在 Pravigard Pac 荟萃分析中，研究者有意采用非比例风险模型，因为“这两种药物中的一种可能在早期（比如第一年）发挥组合疗法的疗效，而另一种则可能在后期接替发挥作用”（S. M. Berry 等人，2004；D. A. Berry，2025）；在 GBM AGILE 和 Precision Promise 中，非比例风险模型也被常规用于总生存期分析。第 Section6.2节所述的安全对数秩检验e过程假设风险成比例，将其扩展到非比例风险情境是一个重要的开放问题。同样，Berry 利用纵向模型根据早期测量预测晚期终点的做法——这一技术在 AWARD-5 中通过在任何患者到达定义性的 12 个月访视前实现无缝阶段转换，“至少节省了一年的开发时间”（D. A. Berry，2025）——展现了一项 e过程方法学尚不具备、但能显著提升其实际应用价值的能力。具有高维干扰参数的复合原假设（例如协变量调整治疗效应）需要采用条件化（这可能牺牲效率），或最坏情况构造（这可能较为保守）。在高维协变量情境中，e值与反应自适应随机化之间的相互作用尚未得到充分理解；在这种情境下，滤过和估计目标都可能随时间变化。更广泛地说，直接学习预测分布的基于分位数的预测方法（Nareklishvili,

Polson 和 Sokolov，2025）提供了一条互补路径，可实现免校准推断，并可能通过共形 e 预测与 E值方法建立联系（Polson、Sokolov 和 Zantedeschi，2026；Datta 等人，2025）。
从监管视角来看，e 值文献与既有监管模板之间的差距正在缩小。2026年FDA贝叶斯指导草案（美国食品药品监督管理局，2026年）虽未提及 e 值，但其成功标准的三路径框架——经 I 类错误率校准、直接后验解释和决策理论——为 E值方法提供了自然的切入点（详见第 9 节）。将结果映射到熟悉的报告指标（始终有效的 $p$值、置信序列），并在统计分析计划中纳入基于模拟的运行特征，仍是获得监管认可的必要步骤；指南要求全面记录运行特征，这同样适用于基于 e 值的设计。

另一项局限在于概念层面：e 值是一种证据尺度，并非完整的决策理论。在确证性试验中，决策受到临床背景、安全性考量、多重性和监管要求的限制。e 值能够实现时间一致的错误控制，但并不能免除明确规定以下事项的必要性：中期观察时允许作出哪些决策、适应性调整会影响哪些人群，以及将报告哪些效应指标。

第10.8 节中的类 RECOVERY 模拟表明，E值方法能够扩展到大型现代试验，但也凸显了一个实际局限：当治疗效应较小（$\delta \approx 0.03$）时，单位观测增长率也相应较小，预期停止时间会超过大多数试验能够招募的患者数量。若能利用通过试验团队数据共享机制获取的患者级数据，对实际的 RECOVERY 试验开展全面的回顾性 e 值分析，就能为比较 e值、组序贯方法和贝叶斯监测提供基准；在该试验中，中期决策对患者护理产生了实质影响（RECOVERY 协作组 2021年）。

## 13 结论

e值和e过程为可选停止和继续情况下的任意时刻有效推断提供了一个原则明确的框架，因此自然适用于适应性临床试验的中期监测。其主要优点包括：无需预先指定停止时刻即可实现明确的时间一致第一类错误控制；在不同观察次数之间（并在一定条件下跨研究）连贯地累积证据；通过下注构造自然处理复合原假设；以及在设计变更是基于既往信息的可预测函数时，能够与自适应决策相兼容。其主要缺点包括：当监测计划固定时，相较于经过最优校准的组序贯界值存在功效差距；效率取决于e过程的选择和调优；处理复杂终点和高维干扰参数还需要额外的方法学研究；此外，当前e值文献与成熟监管模板之间仍存在差距。

针对双臂二分类试验的数值研究给出了具体量化结果：在错误控制方法中，采用固定的20次分析计划时，经校准的组序贯边界检验效能最高（86.1%），校准的贝叶斯规则则更为保守（68.8%），而 e值提供了稳健的任意时刻有效控制（72.3%）。扩展模拟表明，在连续监测下，这一排序会逆转（e值为75%，组序贯方法为10%）；功效差距则因效应量和基线率而异，介于2至29个百分点之间。无效性监测示例证实，e过程能够有效开展无效性监测；类RECOVERY模拟则表明，该方法可扩展至大型试验，同时凸显了小效应带来的样本量代价。这些发现支持采用混合方法，让e值与组序贯方法和贝叶斯方法相互补充，而非取而代之。
对于实践者，建议可简述如下：当试验确实可能需要灵活监测，或根据数据继续、调整试验，且在这些操作下获得可审计的有效性保证是首要要求时，e值最为适用——Berry的平台试验（I-SPY 2、GBM AGILE、REMAP-CAP）已表明，这类情形正日益普遍（D. A. Berry，2025）。当监测计划稳定，且在标准假设下实现面向监管的功效效率至关重要时，传统组序贯设计更为合适。正如Berry在众多具有里程碑意义的试验中所展示的，当实际决策明确关注试验未来成功的概率，且希望由算法驱动入组和分配并实时调整时，贝叶斯预测概率是自然的选择；在这类设计中，如需频率学派操作特征方面的保证，应对阈值进行校准。实践中，三种框架结合使用效果最佳：利用贝叶斯预测概率算法进行实际决策，在预先计划的期中分析中利用组序贯边界开展功效优化的确证性分析，并将e过程作为任意时刻有效的证据账本，无论数据被检查多少次、计划如何变化，或试验是否超出预定期限继续进行，它都能保持有效。

监管环境有利于这种整合。2026年FDA贝叶斯指导草案（美国食品药品监督管理局 2026年）为成功标准提供了三条路径——I类错误校准、直接后验解释，

以及决策理论方法——每条路径都能将 e 值监测纳入其中，作为一种时间一致的证据层。

The `evalinger` R软件包（https://github.com/VadimSokolov/evalinger）及其交互式网络应用（https://sailtargets.shinyapps.io/evalinger/）实现了完整的方法体系，并补充了 Sokolova 和 Sokolov（2026）对适应性试验设计的更广泛论述。该软件包提供 GROW 最优设计校准、实时监测、置信序列、无效性分析、平台试验多重性控制，以及与组序贯方法和贝叶斯规则的直接比较。该网络应用的三个仪表板（设计计算器、监测仪表板和方法比较）可生成美国食品药品监督管理局2026年指导草案要求的运行特征文档（美国食品药品监督管理局2026）：I 类错误率、检验效能曲线、设计先验下的预期样本量，以及对校准参数的敏感性。

Alexander, Brian M., Steffen Ba, Mitchel S. Berger, Donald A. Berry, Webster K. Cavenee, Susan M. Chang, Timothy F. Cloughesy 等。2018年。“胶质母细胞瘤自适应全球创新学习环境：GBM AGILE。”《临床癌症研究》24（4）：737–43。https://doi.org/10.1158/1078-0432.CCR-17-0764。Angus, Derek C., Scott Berry, Roger J. Lewis, Farah Al-Beidh, Yaseen Arabi, Wilma van Bentum-Puijk, Zahra Bhimani 等。2020年。“REMAP-CAP（社区获得性肺炎随机嵌入式多因素自适应平台）研究：研究依据与设计。”《美国胸科学会年鉴》17（7）：879–91。https://doi.org/10.1513/AnnalsATS.202003-192SD。Armitage, Peter。1954年。“预防性和治疗性试验中的序贯检验。”《医学季刊》23（91）：255–74。———。1960年。《序贯医学试验》。牛津：Blackwell Scientific Publications。Baas, Stef, Peter Jacko, and Sofía S. Villar。2025年。“反应自适应临床试验的精确统计分析：一种通用且计算上可行的方法。”https://doi.org/10.48550/arXiv.2407.01055。Barker, Anna D., Carrie C. Sigman, Gary J. Kelloff, Nola M. Hylton, Donald A. Berry, and Laura J. Esserman。2009年。“I-SPY 2：新辅助化疗背景下的自适应乳腺癌试验设计。”《临床药理学与治疗学》86（1）：97–100。https://doi.org/10.1038/clpt.2009.68。Berry, Donald A.。1989年。“临床试验中累积数据的监测。”《生物统计学》45（4）：1197–1211。https://doi.org/10.2307/2531771。———。2004年。“贝叶斯统计学与临床试验的效率与伦理。”《统计科学》19（1）：175–87。https://doi.org/10.1214/088342304000000044。———。2025年。“自适应贝叶斯临床试验：临床研究的过去、现在与未来。”《临床医学杂志》14（15）：5267。https://doi.org/10.3390/jcm14155267。Berry, Donald A., Shobha Dhadda, Michio Kanekiyo, Dayan Li, Chad J. Swanson, Michael Irizarry, Larry D. Kramer, and Scott M. Berry。2023年。“仑卡奈单抗治疗早期阿尔茨海默病患者：一项2b期剂量探索随机临床试验的贝叶斯分析。”《JAMA 网络开放》6（3）：e237230。https://doi.org/10.1001/jamane tworkopen.2023.7230。Berry, Donald A., and Bert Fristedt。1985年。《老虎机问题：实验的序贯分配》。伦敦：Chapman; Hall。Berry, Scott M., Donald A. Berry, Karthik Natarajan, Chyi-Shya Lin, Charles H. Hennekens, and Robert Belder。2004年。“非比例风险下的贝叶斯生存分析：普伐他汀-阿司匹林的荟萃分析。”《美国统计学会杂志》99（465）：36–44。https://doi.org/10.1198/016214504000000106。Berry, Scott M., Bradley P. Carlin, J. Jack Lee, and Peter Müller。2010年。《临床试验的贝叶斯自适应方法》。Chapman & Hall/CRC 生物统计学系列。佛罗里达州博卡拉顿：CRC 出版社。https://doi.org/10.1201/EBK1439825488。Datta, Jyotishka, Nicholas G. Polson, Vadim Sokolov, and Daniel Zantedeschi。2025年。“保形预测 = 贝叶斯？”https://doi.org/10.48550/arXiv.2512.23308。Evans, Michael, and Hadas Moshonov。2006年。“检查先验与数据冲突。”《贝叶斯分析》1（4）：893–914。https://doi.org/10.1214/06-BA129。Geiger, Mary Jane, Zachary Skrivanek, Brenda L. Gaydos, Jenny Y. Chien, Scott M. Berry, and Donald A. Berry。2012年。“长效胰高血糖素样肽-1类似物（度拉糖肽）的自适应剂量探索无缝2/3期研究：试验设计与基线特征。”《糖尿病科学与技术杂志》6（6）：1319–27。https://doi.org/10.1177/193229681200600610。Grünwald, Peter, Rianne de Heide, and Wouter Koolen。2023年。“安全检验。”https://doi.org/10.48550/arXiv.1906.07 801。Grünwald, Peter, Alexander Ly, Muriel Perez-Ortiz, and Judith ter Schure。2021年。“安全对数秩检验：可选停止、继续及先验设定错误下的错误控制。”载于《2021年 AAAI 生存预测春季研讨会：算法、挑战与应用论文集》，146：107–17。《机器学习研究论文集》。PMLR。https://proceedings.mlr.press/v146/grunwald21a.html。

Howard, Steven R., Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon。2021年。“时间一致、非参数、非渐近置信序列。”《统计年鉴》49（2）：1055–80。https://doi.org/10.1214/20-AOS1991。Irony, Telba Z., and Gregory Campbell。2011年。“医疗器械临床试验中的贝叶斯方法：监管环境中的实例讨论。”《生物制药统计学杂志》21（1）：81–115。https://doi.org/10.1080/10543406.2011.536168。Irony, Telba Z., Gene Pennello, Jozef Engel, Honghong Gao, Zhiwen Meng, Chunyan Chen, Fang Liu, Yunling Xu, and Kimberly L. Price。2023年。“医疗器械贝叶斯统计学：2010年以来的进展。”《治疗创新与监管科学》57（3）：401–13。https://doi.org/10.1007/s43441-022-00495-w。Jennison, Christopher, and Bruce W. Turnbull。2000年。《组序贯方法及其在临床试验中的应用》。Chapman & Hall/CRC。Lan, K. K. Gordon, and David L. DeMets。1983年。“临床试验的离散序贯界限。”《生物统计学》70（3）：659–63。https://doi.org/10.1093/biomet/70.3.659。Martin, Ryan。2025年。“正则化e过程：通过基于知识的效率提升实现任意时刻有效推断。”https://doi.org/10.48550/arXiv.2410.01427。Meurer, William J., Roger J. Lewis, and Donald A. Berry。2012年。“适应性临床试验：对治疗性误解的部分补救？”JAMA 307（22）：2377–78。https://doi.org/10.1001/jama.2012.4966。Morita, Satoshi, Peter F. Thall, and Peter Müller。2008年。“确定参数先验的有效样本量。”《生物统计学》64（2）：595–602。https://doi.org/10.1111/j.1541-0420.2007.00888.x。Muss, Hyman B., Donald A. Berry, Constance T. Cirrincione, Maria Theodoulou, Ann M. Mauer, Alice B. Kornblith, Ann H. Partridge 等。2009年。“老年早期乳腺癌女性的辅助化疗。”《新英格兰医学杂志》360（20）：2055–65。https://doi.org/10.1056/NEJMoa0810266。Nareklishvili, Maria, Nick Polson, and Vadim Sokolov。2025年。“生成式分位数贝叶斯预测。”https://doi.org/10.48550/arXiv.2510.21784。Novick, Melvin R., and James E. Grizzle。1965年。“临床试验数据分析的贝叶斯方法。”《美国统计学会杂志》60（309）：81–96。https://doi.org/10.1080/01621459.1965.10480776。O’Brien, Peter C., and Thomas R. Fleming。1979年。“临床试验的多重检验程序。”《生物统计学》35（3）：549–56。https://doi.org/10.2307/2530245。Park, John W., Minetta C. Liu, Douglas Yee, Christina Yau, Laura J. van ’t Veer, W. Fraser Symmans, Melissa Paoloni 等。2016年。“早期乳腺癌中奈拉替尼的自适应随机分配。”《新英格兰医学杂志》375（1）：11–22。https://doi.org/10.1056/NEJMoa1513750。Pennello, Gene, and Laura Thompson。2007年。“审评贝叶斯医疗器械试验的经验。”《生物制药统计学杂志》18（1）：81–115。https://doi.org/10.1080/10543400701668274。Pocock, Stuart J.。1977年。“临床试验设计与分析中的组序贯方法。”《生物统计学》64（2）：191–99。https://doi.org/10.1093/biomet/64.2.191。Polson, Nick, Vadim Sokolov, and Daniel Zantedeschi。2026年。“贝叶斯、E 值与检验。”https://doi.org/10.48550/arXiv.2602.04146。Ramdas, Aaditya, Peter Grünwald, Vladimir Vovk, and Glenn Shafer。2023年。“博弈论统计学与安全的任意时刻有效推断。”《统计科学》38（4）：576–601。https://doi.org/10.1214/23-STS894。RECOVERY 协作组。2021年。“地塞米松用于住院 COVID-19 患者。”《新英格兰医学杂志》384（8）：693–704。https://doi.org/10.1056/NEJMoa2021436。Robbins, Herbert。1952年。“实验序贯设计的若干方面。”《美国数学会通报》58（5）：527–35。https://doi.org/10.1090/S0002-9904-1952-09620-8。Saville, Benjamin, Donald A. Berry, Natalie S. Berry, Kert Viele, and Scott M. Berry。2022年。“贝叶斯时间机器：处理多臂平台试验中的时间漂移。”《临床试验实践》19（5）：490–501。https://doi.org/10.1177/17407745221112013。Schure, Judith ter, and Alexander Ly。2022年。“ALL-IN 元分析：为活的系统综述注入活力。”F1000Research 11：549。https://doi.org/10.12688/f1000research.74223.2。Shafer, Glenn。2021年。“押注式检验：一种统计与科学传播策略。”《皇家统计学会杂志：A辑（社会统计学）》184（2）：407–31。https://doi.org/10.1111/rssa.12647。Shafer, Glenn, and Vladimir Vovk。2019年。《概率与金融的博弈论基础》。新泽西州霍博肯：Wiley。Sokolova, Alexandra, and Vadim Sokolov。2026年。《临床试验实践：从基础到人工智能代理》。https://vsokolov.quarto.pub/clinos/。Spiegelhalter, David J., Keith R. Abrams, and Jonathan P. Myles。2004年。《临床试验与医疗保健评估的贝叶斯方法》。第13卷。奇切斯特：John Wiley & Sons。Thompson, William R.。1933年。“根据两个样本所提供的证据，一个未知概率大于另一个的似然。”《生物统计学》25（3–4）：285–94。https://doi.org/10.1093/biomet/25.3-4.285。

美国食品药品监督管理局。2019年。《药物和生物制品临床试验的适应性设计：行业指南》。https://www.fda.gov/media/78495/download。———。2025年。《医疗器械临床试验中的贝叶斯统计：行业及 FDA 工作人员指南》。https://www.fda.gov/media/190505/download。———。2026年。《药物和生物制品临床试验中贝叶斯方法学的应用：行业指南（指南草案）》。https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs。Kert Viele。2026年。《FDA 2026年贝叶斯指南草案指南》。 https://www.berryconsultants.com/resource/guide-to-the-draft-fda-bayesian-guidance-2026。Jean Ville。1939年。《集体概念批判研究》。巴黎：Gauthier-Villars。Vladimir Vovk 和 Ruodu Wang。2021年。《e 值：校准、组合与应用》。《统计年鉴》49（3）：1736–54。https://doi.org/10.1214/20-AOS2020。Abraham Wald。1947年。《序贯分析》。纽约：John Wiley & Sons。Ruodu Wang 和 Aaditya Ramdas。2022年。《基于 e 值的错误发现率控制》。《皇家统计学会会刊：*B*辑（统计方法学）》84（3）：822–52。https://doi.org/10.1111/rssb.12489。Janet Woodcock 和 Lisa M. LaVange。2017年。《用于研究多种疗法、多种疾病或两者兼有的主方案》。《新英格兰医学杂志》377（1）：62–70。https://doi.org/10.1056/NEJMra1510062。