预印本。审稿中。

SECRETS：使用合成干预的受试者高效临床随机对照试验

Sayeri Lala SLALA@PRINCETON.EDU
电气与计算机工程系，普林斯顿大学，美国新泽西州普林斯顿市，邮编 *08544*
Niraj K. Jha JHA@PRINCETON.EDU
电气与计算机工程系，普林斯顿大学，美国新泽西州普林斯顿市，邮编 *08544*

arXiv:2305.05078v1 [eess.SP] 8 May 2023

摘要

随机对照试验（RCT）是估计医学干预平均治疗效应（ATE）的金标准，但需要数百至数千名受试者，因此成本高昂且难以实施。交叉试验通过测量每位受试者的治疗效应来降低样本量要求，但仅适用于慢性疾病以及效应能够迅速消退的干预。另一种方法是用前瞻性研究或既往随机对照试验的外部数据替代或补充随机对照试验收集的数据，但外部数据或补充数据中的混杂因素可能影响结果。我们提出模拟交叉试验，以克服其实际应用中的局限，同时发挥其优势。我们提出了一种新型框架 SECRETS：该框架首次利用一种名为合成干预的最先进反事实估计算法，在不使用任何外部数据的情况下，估计随机对照试验研究中每位患者的个体治疗效应（ITE）。此外，该框架还采用一种新的假设检验策略，根据 ITE 估计值判断治疗是否具有临床意义上的平均处理效应。我们表明，SECRETS 能够在保持显著性水平相当的同时提高随机对照试验的检验效能；具体而言，在三项真实世界临床随机对照试验（III期试验）中，与基线方法相比，SECRETS 将检验效能提高了 **6-54%**（平均值：21.5%，标准差：15.8%）。

关键词— 临床随机对照试验、反事实估计、假设检验、样本效率、合成干预。

1. 引言

随机对照试验（RCT）是估计药物、治疗或其他医学干预的人群层面平均治疗效应（ATE）的黄金标准方法 [1, 2]。通常在III期临床研究中使用，以确定治疗的效果 [1, 3]。典型的随机对照试验，即双臂平行组优效性试验 [1], ，通过估计平均治疗效应并判断估计结果是否具有统计学显著性，评估某种治疗是否优于标准治疗或安慰剂。为估计平均治疗效应，随机对照试验首先招募能够代表目标人群的受试者，并将其随机分配至对照组和治疗组，受试者分别接受标准治疗/安慰剂和目标治疗。然后，

。本文已提交 IEEE，等待其考虑是否发表。。© 20xx IEEE。允许个人使用本材料。任何其他用途均须获得 IEEE 的许可，无论是在当前还是未来的任何媒体中，包括为广告或促销目的而重印或重新发布本材料、创作新的合集、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护部分。

它在整个研究期间并行监测两组中预先指定的健康指标，评估每个治疗组下由该健康指标定义的平均治疗结局，并将两者之差报告为平均处理效应。随机分配对于准确估计平均处理效应至关重要，因为它通过在对照组和治疗组之间形成更具可比性的队列来消除混杂因素，并允许使用标准方法的统计假设检验，即双侧检验，以所需的准确度检测目标平均处理效应 [1]。然而，由于响应中的受试者间变异，随机对照试验通常需要数百至数千名受试者，才能以高准确度检测出具有临床意义的平均处理效应 [3, 4], ，这使其成本高昂（2021年每位参与者平均至少需42,000美元 [5]）且难以实施，而招募通常是“最困难的任务” [1]。

提高随机对照试验受试者效率的一种方法是采用交叉设计。与平行随机对照试验相比，该设计通过降低平均处理效应估计值的方差来提高检验效能 [6]。具体而言，它会针对每位受试者测量个体处理效应（即个体在处理组和对照条件下响应的差异），再对所有个体处理效应取平均值。由于观测值数量翻倍（每位受试者两次），且预期结局指标的方差会因消除患者间变异而降低，因此平均处理效应估计值的方差低于平行随机对照试验 [1, 6]。尽管交叉设计能够降低相对于平行随机对照试验的抽样复杂度 [6, 7], ，但它也面临一些限制，因而无法广泛应用。为了解这些限制，我们来考察最简单的版本，即两期交叉设计。该设计通过为每位患者分配一个随机治疗序列来比较两种治疗（例如，标准治疗/安慰剂与目标治疗）：患者在第一期接受标准治疗或目标治疗，并在第二期接受另一种治疗；随机分配的序列用于排除时间效应。重要的是，两期之间会插入一个洗脱期，以消除第一期所施治疗产生的任何残留效应。由于需要设置洗脱期，交叉设计可研究的疾病-干预组合类型受到限制，因此仅适用于治疗效应能够迅速消退的慢性疾病 [1, 6]。

另一种方法并非试图从小规模随机样本中估计平均处理效应，而是利用从既往临床试验、电子健康记录、患者登记库等来源获得的大型外部数据集，来降低平均处理效应估计值的方差。 [8] 之所以使用这些数据集来补充或替代现有随机对照试验数据，是因为它们获取成本可能更低、也更容易获得， [1, 5] 并且与随机对照试验数据相比，更能代表目标人群。 [9]然而，由于整合数据集不再具有随机性，因此从中估计平均处理效应十分困难。具体而言，外部数据可能在多个方面不同于我们关注的随机对照试验数据，例如受试者特征、医疗方案等；合并这两个数据集可能引入混杂因素， [1, 8], 即同时影响治疗反应和治疗分配的变量，从而产生不可比的对照组和治疗组，进而无法准确估计平均处理效应。 [2]因此，需要仔细筛选外部对照数据，最好选自数据收集方法、研究终点和研究人群具有可比性的历史随机对照试验，以减少其与目标随机对照试验之间的差异。 [8]为了在存在混杂因素的情况下进行学习，人们已经开发出从观察性数据中估计平均处理效应的方法，包括通过重新加权构建新的可比对照组和治疗组， [8]以及在计算平均处理效应时通过分层对混杂因素进行条件化。 [2]此外，人们还开发了反事实估计算法来估计个体处理效应 [2]；然而，这些算法仅估计治疗组的个体处理效应，这很可能是因为存在大量历史对照数据；因此，估计得到的平均处理效应仅基于治疗组的个体处理效应估计值。 [10, 11]但是，所有这些从观察性数据中学习的方法，仅在不存在未观测混杂因素，且对照组与治疗组的支撑集（协变量分布）存在重叠时才有效；而这些条件不太可能成立，因为要验证不存在未观测混杂因素需要领域专业知识，并且在高维数据集上进行操作会减少重叠。 [9]由于存在这些条件限制，试图从

观察性数据中估计平均处理效应，无论是与现有随机对照试验数据结合，还是单独使用，都需要比随机对照试验更大的样本量。 [5, 10, 12]

我们不使用外部对照数据集，也不面对数据整理和分析方面的挑战，而是提出模拟交叉设计，以利用方差缩减原理，同时避免其实际限制。给定一个随机对照试验数据集，我们的框架称为 SECRETS（使用合成干预的受试者高效临床随机对照试验，**S**ubject-**E**fficient **C**linical **R**andomiz**E**d Controlled **T**rials using **S**ynthetic Intervention），首先利用最先进的反事实估计算法——合成干预（SI） [13], ，仅根据所有受试者同时期的对照组和治疗组，估计每位受试者的个体处理效应；该算法能够估计控制和治疗条件下的反事实；然后对个体处理效应取平均值，以估计平均处理效应，并应用一种新颖的假设检验算法控制第一类错误率，从而高精度地检测目标平均处理效应。

我们的工作在多个方面有别于相关研究。一些研究在没有外部对照数据的情况下进行个体处理效应估计，但它们利用估计得到的个体处理效应，在随机对照试验研究人群中识别治疗效应增强的亚组 [14, 15], ，以支持随机对照试验的主要发现[1]。相比之下，我们利用个体处理效应估计在给定样本量下提高检验效能，从而减少回答随机对照试验主要问题所需的样本量，即研究人群中是否存在目标平均处理效应；因此，先前这些工作所解决的问题与我们所研究的问题是正交的。此外，这些工作使用虚拟孪生（VT）算法 [14]来估计用于计算个体处理效应的反事实。VT 算法基于协变量拟合回归模型（例如随机森林），以预测患者在接受不同于其实际治疗的治疗时的反事实响应。相比之下，我们的框架使用 SI，这是一种非参数算法，利用所有被分配至某项干预的患者的观测响应，估计患者在给定干预下的反事实响应。

由于 SI 是与我们最相关的研究，我们将说明本研究与其有何不同。首先，尽管 SI 已被评估用于实现数据高效的随机对照试验 [13], ，但它的用途是减少每个个体所需实施的干预 次数。由于标准临床随机对照试验仅将每位受试者分配到一种干预，我们则展示了如何利用 SI 减少分配到每种干预的受试者数量。此外，SI 框架对因果治疗效应的定义，以及用于评估数据效率的指标，都不同于临床随机对照试验中的定义和指标。SI 将因果治疗效应定义为给定干预下响应的时间平均值，而临床随机对照试验则将其定义为平均处理效应，或处理组与对照干预下响应的差异。SI 使用估计因果参数的平方误差指标进行衡量；而我们则衡量检测目标平均处理效应时所达到的显著性水平和检验效能，因为临床随机对照试验会依据这些数值确定样本量。最后，用于评估 SI 的电商随机对照试验与临床随机对照试验的数据特征并不相似，这可能会影响其研究结果的普适性；尤其是，该电商研究的干预前数据涵盖八个时间点，可能提升了性能 [16], ，且研究单位是客户群体，其响应是对个体响应取平均值后得到的，因此潜在的供体噪声更小。相比之下，在我们的框架中，临床随机对照试验通常只有一次干预前测量或基线测量，研究单位则是个体。

鉴于我们的工作与以往及相关工作的差异，我们将贡献总结如下：

1. 我们提出了一种新颖的框架 SECRETS，该框架首次在**不依赖任何外部数据**的情况下，相较于标准随机对照试验（RCT）提高了检验效能，进而提升了受试者效率。具体而言，首先，该框架利用一种最先进的反事实估计算法 SI（能够在对照条件和处理条件下进行反事实估计），估计同期对照组和治疗组中所有患者 的个体处理效应。随后，利用估计得到的个体处理效应降低平均处理效应估计值的方差，并针对 SI 估计的个体处理效应违反标准假设检验策略所要求的假设这一情况，实施一种新颖的假设检验程序，以控制第一类错误率。

2. 我们在三个真实世界的临床随机对照试验数据集上验证了 SECRETS，结果表明，与基线方法相比，该方法可将检验效能提高 6%–54%（平均值：21.5%，标准差：15.8%），从而使对照组和治疗组达到典型目标统计工作点（即显著性水平为 5%、检验效能为 80%）所需的样本量减少 25%–76%，即减少 10–3,957 个样本。

3. 我们通过消融研究说明，反事实估计算法和假设检验策略对于 SECRETS 的成功至关重要。

本文其余部分安排如下。第 2 节介绍与我们提出的框架相关的背景知识。第 3 节描述我们提出的框架。第 4 节说明我们如何评估性能。第 5 节展示结果，第 6 节讨论其意义，第 7 节总结全文。

2. 背景

本节回顾理解 SECRETS 所需的相关概念，即反事实估计算法和假设检验方案。

2.1 反事实结果估计的合成干预

SI 是一种最先进的反事实估计算法，对于给定个体（例如患者），它可以估计该个体在干预期内未接受的其他干预下的反事实结果轨迹 [13]。为此，首先假设存在一个接受过各种干预（包括对照干预）的供体个体群体，并且在干预前期，每个个体都被分配至对照组。然后，对于在某种干预下观测到的给定目标单元 j，SI 通过对接受干预的供体个体的干预后结果进行加权，计算该目标单元在不同干预 k 下的反事实结果 k。它使用主成分回归学习供体个体的权重，使其干预前结果的加权和能够最佳地预测目标单元的干预前结果。与匹配和重新加权等其他方法相比，SI 的有效性 [2]，源于其能够在存在未观测混杂因素时预测反事实 [13]。

2.2 假设检验

假设检验是一种统计程序，用于确定感兴趣参数 μ（例如，某种药物的平均处理效应）是否为某个指定的非零值。连续参数的标准设置会考虑两个假设：零假设，即认为 $\mu = \mu_0 = 0$；备择假设，即认为 $\mu = \mu_1 > 0$ （在我们关注其大小的双侧检验中） [1]。目标是设计一种检验，在限制第一类错误或显著性水平 α（当零假设为真时识别出 $\mu = \mu_1$ 的概率）的同时，使检验效能 $1-\beta$（当备择假设为真时识别出 $\mu = \mu_1$ 的概率）最大化。该检验由根据观测值计算出的检验统计量，以及由用于控制显著性水平的临界值确定的拒绝域构成；如果统计量落入拒绝域，则拒绝零假设，否则无法拒绝零假设。

检验策略会根据对零假设下观测值分布的假设，采用不同方式选择检验统计量和拒绝域，以提高检验效能。例如，针对正态分布均值的双侧 t 检验假设观测值独立同分布于一个方差未知的正态分布，并将样本均值除以样本方差的平方根后所得的值设为检验统计量，将 Student t 分布（具有适当自由度）的第（$1-\alpha/2$）百分位数设为临界值，其中

自由度），α 为目标显著性水平。当数据分布未知时，自助法检验会对数据进行重抽样，并根据 α [17, 18] 设定临界值。

然而，这些检验策略假设数据样本中的观测值相互独立。当独立性不成立时，应用这些检验会导致显著性水平高于预期。为解决这一问题，已有研究提出了几种改进方法 [19]。一种方法试图通过考虑数据依赖性所引起的协方差项，校正统计量中使用的样本方差估计；然而，实证结果表明，当相关结构是根据数据估计时，这种策略会产生较大的方差估计 [19]。另一种方法试图通过拟合自回归模型来模拟数据依赖性，并从中抽样以构建用于检验的置信区间，从而改进相关结构的估计；然而，在样本量较大时，这种方法的效率可能较低 [19]。受限随机化方法不直接估计相关结构，而是构建能够考虑依赖结构的零分布；然而，这些方法假设数据具有特殊的依赖结构 [20]。

3. 方法

在回顾必要背景后，我们现在介绍我们的框架 SECRETS。首先，我们概述该框架，并解释其如何降低标准方法随机对照试验（即双臂平行组优效性试验）的抽样复杂度。然后，我们将详细描述该框架的每个步骤。

概述

SECRETS 框架如图 1 所示，在给定样本量下提高检验效能，从而减少达到目标检验效能 $1-\beta_{target}$、目标 ATE $\mu_{1,\ target}$和显著性水平 α_{target}所需的样本量，相较于标准方法 RCT。它模拟交叉设计，以降低平均处理效应估计值的方差。给定从已完成的随机对照试验中收集的数据，即观察到的对照组数据 X_{ctrl} 和处理组数据 X_{treat}、SI 调优参数，以及一个根据患者在某种干预下的响应轨迹计算相关健康结局的函数 $get\ outcome$，SECRETS 首先估计每个组中所有患者的个体处理效应，即，

Y_{ctrl} 和 Y_{treat}。随后，它使用对照组数据、合并的个体处理效应、SI 调优参数、函数 获取结果、从零分布中生成的样本数（T）以及用于检验的其他参数来实施并运行双侧假设检验（输入说明请参见附录 A、算法 1）。接下来，我们将详细介绍 SECRETS 的每个步骤。

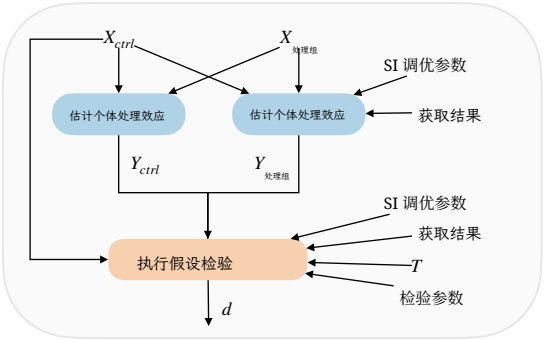


图 1：SECRETS 框架的流程图。请注意，对 估计个体处理效应 的两次调用都接收 SI 调优参数和 获取结局，-但为简洁起见，我们省略了这些参数。

步骤 1：估计个体处理效应

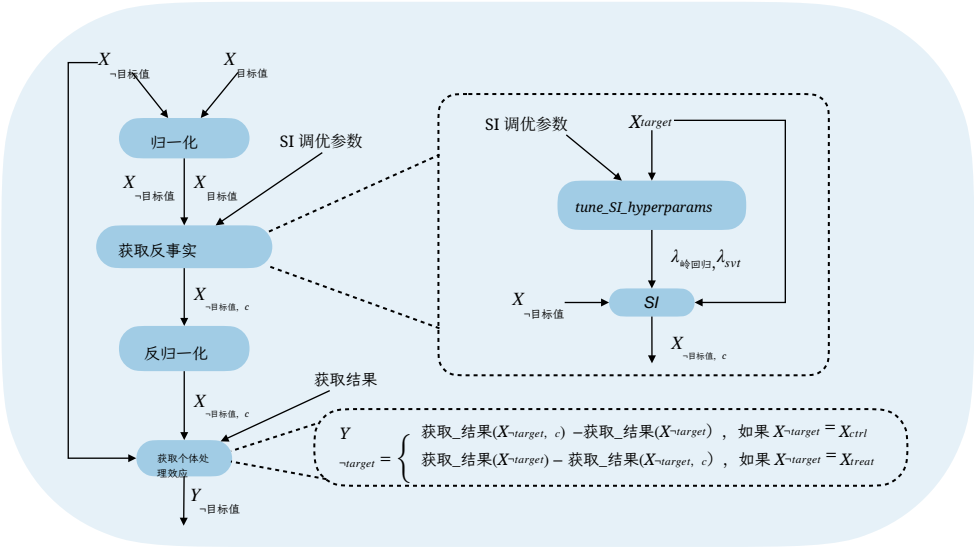


图 2：子程序“估计个体处理效应”的流程图。

为了估计个体处理效应 Y_{ctrl} 和 Y_{treat}，SECRETS 使用图 2 所示的子程序 估计个体处理效应。给定接受目标干预的组（X_{target}）的数据，以及“未暴露”组（$X_{\neg target}$）的数据（即接受不同干预的组），估计个体处理效应 首先执行最小-最大归一化，并调用 获取反事实 来插补未暴露组在目标干预下的响应（$X_{\neg target,c}$）。为此，获取反事实 先使用 SI 调优其超参数（λ_{ridge} 和 λ_{svt}）：利用子程序 调优 *SI* 超参数 和 SI 调优参数在 X_{target} 上进行调优（见附录 A、算法 2）；然后运行 $\overline{SI}$，将 X_{target} 作为供体组，并使用调优后的超参数，将 $X_{\neg target}$ 中的每个单元作为目标单元（见附录 A、算法 3）。对 $X_{\neg target,c}$执行反归一化后，估计个体处理效应 调用 获取个体处理效应，该函数分别在 $X_{\neg target,c}$ 和 $X_{\neg target}$上（向量化地）调用 获取结果，计算目标干预与观测到的干预下结果之间的差值。具体而言，SECRETS 通过将 $X_{\neg target}$设为 X_{ctrl}、将 X_{target} 设为 X_{treat} 来计算 Y_{ctrl}，反过来设置则可计算 Y_{treat}。

步骤 2：进行数据驱动假设检验

计算完个体处理效应（ITEs）后，SECRETS 会执行假设检验，利用 ITEs 确定是否拒绝零假设。标准的双侧单样本检验并不适用，因为估计的 ITEs 彼此并非独立，这是由于 SI 使用相同的供体池，即处理组（对照组），来估计对照组（处理组）中每个目标单元的反事实。此外，现有的处理相关样本的方法要么在实际应用中表现不佳，要么对依赖结构作出了不适用于 ITE 估计值的假设。因此，SECRETS 采用了一种新的假设检验程序执行假设检验，如图 3 所示，该程序能够考虑 ITEs 之间的依赖关系，同时避免上述局限性。

具体而言，运行假设检验实现了双侧单样本t检验所用的临界值检验程序，但会利用数据来调优临界值 t_c。为调优 t，运行假设检验首先通过从检验统计量的零分布中抽样 T 次来近似其零分布（$\hat{s}_{null}$），每次抽样使用 $sample\ null$。为生成第 i 个样本，零假设抽样首先从以下数据中进行有放回抽样

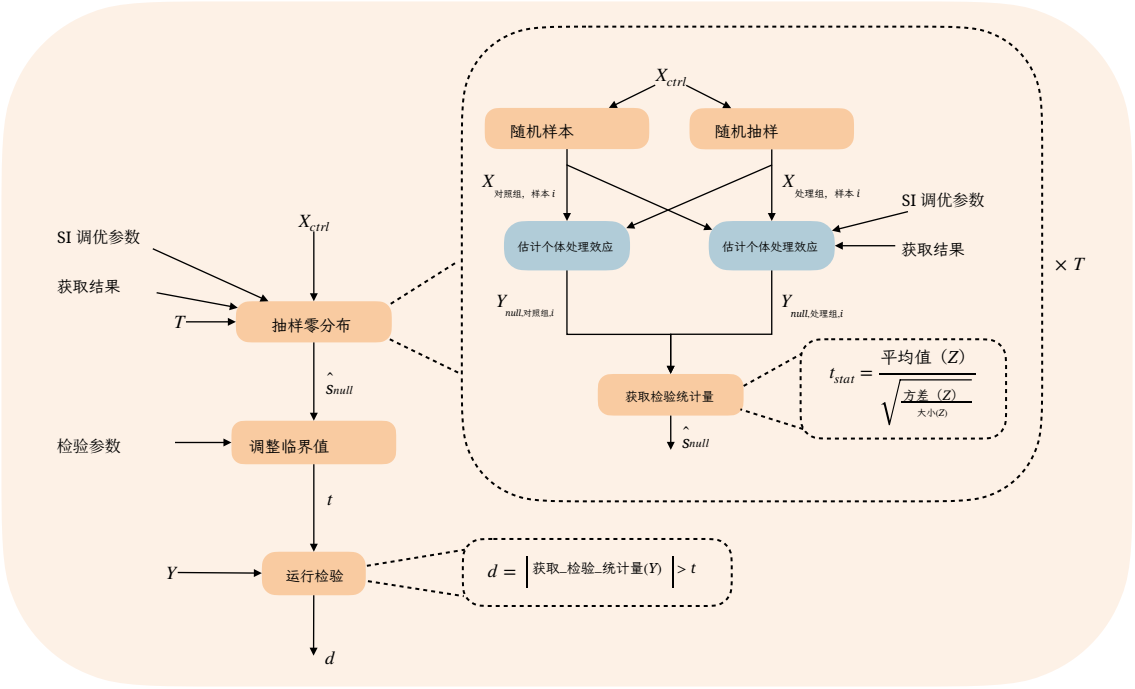


图 3：运行假设检验子程序的流程图。请注意，在零假设抽样中，两次调用估计个体处理效应都会使用 SI 调优参数和获取结果函数，但为简洁起见，我们省略了参数。

原始对照组数据（X_{ctrl}），使用随机样本构造对照组和治疗组，使其受试者人数与原始对照组和治疗组相同，且响应具有可比性（$X_{ctrl,\ sample_i}$ 和 $X_{treat,\ sample_i}$）。随后，与步骤 1 一样，它运行估计个体处理效应（使用 SI 调优参数和函数获取结果），计算构造出的对照组和治疗组的个体处理效应。由于构造出的对照组和治疗组接受了相同的干预（即对照条件），相应的个体处理效应 $Y_{null,\ ctrl,\ i}$ 和 $Y_{null,\ treat,\ i}$均为零分布的样本。给定合并后的个体处理效应集合，获取检验统计量使用单样本t检验的检验统计量公式，计算所得检验统计量 $\hat{}s_{null,\ i}$。

随后，运行假设检验 使用调节临界值来调节检验的临界值。该方法利用 $\hat{s}_{null}$ 和一组检验参数（包括目标显著性水平 α_{target} 和搜索范围参数），以类似二分搜索的方式找到达到目标显著性水平的临界值 t（见附录 A、算法 4）。

最后，运行假设检验 通过根据步骤 1 中原始对照组和治疗组（Y）得出的个体处理效应计算检验统计量，并将其绝对值与调节后的临界值 t进行比较，来执行双侧检验；若 $d = 1$，则拒绝零假设，否则不拒绝零假设。

4. 性能评估设置

本节介绍用于评估 SECRETS 的实验设置。首先，我们介绍性能指标。然后，介绍用于理解该框架的基线方法和消融实验。最后，介绍用于开展实验的数据集。

指标

为评估样本效率，我们比较每组给定样本量为 n_a 名受试者（即对照组和治疗组）且目标显著性水平为 α_{target}时得到的检验效能。我们还报告达到期望检验效能 $1-\beta_{target}$所需的样本量。为衡量检验效能 $1-\beta$和显著性水平 α，我们采用 [7], 的方法：分别在备择设定和零假设设定下模拟大量试验，并计算检验程序返回拒绝的试验所占百分比。首先，我们将备择设定下的目标平均处理效应定义为在完整随机对照试验数据集上测得的平均处理效应，即$\mu = \mu_{1,\ target}$，并将 $\mu = \mu_0 = 0$ 定义为零假设设定下的平均处理效应。然后，为衡量每组样本量为 n_a 名受试者时的检验效能 $1-\beta$，我们运行 L 次试验；通过分别从原始随机对照试验的对照组和治疗组中有放回地抽取 n_a 名受试者，构建新的对照组和治疗组，以模拟备择设定下的试验。同样，我们通过 L 次试验衡量显著性水平 α ： 模拟零假设设定，分别从原始随机对照试验的对照组中有放回地抽取 n_a 名受试者，构建对照组和治疗组。

为了理解在备择设定和零假设设定下，估计的平均处理效应（即样本统计量）的分布（均值和方差）变化如何导致检验效能变化，我们将双侧假设检验的检验效能公式 [17], （即式（1））视为一个模型。

$$1-\hat{\beta} \approx \Phi\left[-z_{1-\alpha_{target}/2} + \frac{|\hat{\mu}_0 - \hat{\mu}_1|}{\hat{\sigma}}\right] \tag{1}$$

该公式指出，如果样本统计量的底层分布为正态分布，那么双侧z检验的检验效能 $1-\beta$ 近似等于累积分布函数 Φ 在临界值 $z_{1-\alpha_{target}/2}$ 的相反数处的取值；也就是说，它是一个标准正态分布经常数偏移后的 $(1-\alpha_{target}/2)\%$ 分位数，我们将此常数称为位移项。位移项 由式（2）给出，其中参数 $\hat{}\mu_0$ 和 $\hat{}\mu_1$ 分别对应于零假设设定和备择设定下样本统计量分布的均值，参数 $\hat{}\sigma$ 则为这些分布的标准差。该公式意味着，对于给定的 α_{target}，随着位移项 增大，检验效能也会提高；当均值之间的差距增大和/或分布的方差减小时，就会出现这种情况。因此，如果该公式能准确估计检验效能，我们便可以根据分布的均值和方差差异来解释检验效能的差异。

$$shift\ term = \frac{|\hat{\mu}_0 - \hat{\mu}_1|}{\hat{\sigma}} \tag{2}$$

在我们的分析中，$\hat{}\mu_0$ 和$\hat{}\mu_1$ 分别对应于零假设和备择设定下平均处理效应估计值分布的均值（根据试验得出）。我们通过对零假设和备择设定下平均处理效应估计值分布的标准差取平均，来估计$\hat{}\sigma$ ；同时还报告该估计值的标准差（标准差），以表明这些分布的标准差具有可比性。为评估模型是否成立，我们报告模型在给定 α_{target}、$\hat{}\mu_0$、$\hat{}\mu_1$和$\hat{}\sigma$时对检验效能 $1-\beta$ 的估计。由于小样本量下平均处理效应估计值的非正态性以及检验程序的差异，该方程可能无法准确预测测得的检验效能 $1-\beta$，但我们仍可用它来比较方法，并定性评估分布差异如何解释检验效能的差异。

我们还评估了估计的平均处理效应的质量，因为它们可能具有参考价值 [1]。我们通过测量零假设和备择设定下估计的平均处理效应估计值分布的均值与相应真实值之间的差异，来计算平均处理效应估计值的误差；这些真实值分别为 $\hat{}\mu_0 - \mu_0$ and $\hat{}\mu_1 - \mu_{1,target}$。

基线

我们介绍用于对 SECRETS 进行基准比较的基线方法。我们将该基线方法称为标准方法，它是双臂平行组优效性随机对照试验用于确定处理效果的方法。其流程图如图 4 所示。首先，给定对照组数据和处理组数据，它使用 $get\ outcome$（向量化）计算相应的结局数据 O_{ctrl} 和 O_{treat}。随后，它使用结局数据和期望的显著性水平 α_{target}，对独立样本进行方差不等的双样本t检验[17]。

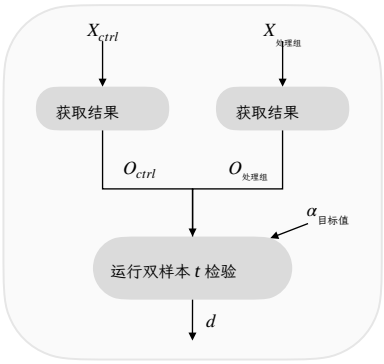


图 4：标准方法 基线的流程图。

标准方法估算每组所需的样本量 n_a，以在目标 ATE 为 μ_1 时达到检验效能 $1-\beta_{target}$，$target$ 且显著性水平为 α_{target}；该估算使用基于双侧 z 检验的方程 [17]，即式（3）。其中，z_m 表示标准正态分布的第 m 个百分位数，σ^2_{ctrl} 和σ^2_{treat} 分别表示对照组和治疗组结局指标的方差，通常根据既往研究估算 [1]。

$$n_a = \frac{(\sigma^2_{ctrl} + \sigma^2_{treat})(z_{1-\alpha_{target}/2} + z_{1-\beta_{target}})^2}{(\mu_{1,target} - \mu_0)^2} \tag{3}$$

在我们的实验中，我们使用完整数据集估计 σ^2_{ctrl} 和 σ^2_{treat}，以获得准确的样本量估计。然而，由于该方程对于基于小数据集的估计不成立 [1, 17]，在此类情况下，我们会增加样本量，直到测得的检验效能 $1-\beta$ 与 $1-\beta_{target}$匹配。

消融研究

在本节中，我们将介绍用于评估 SECRETS 各个基础步骤有效性的消融研究。

首先，我们通过将 SECRETS 的反事实估计算法 SI 替换为 VT 算法Estimate Ites [14], 进行消融实验。VT 算法使用基于协变量拟合的回归模型来估计反事实结局。我们开展这项消融实验，以评估 SI 的优势：SI 学习的是患者特异性模型，而非共享模型。具体而言，我们分别使用来自对照组和治疗组的所有患者数据，拟合两个共享模型：一个用于预测对照条件下的响应，另一个用于预测治疗条件下的响应；然后分别使用这两个模型，预测对照组和治疗组中每位患者的干预后反事实响应。与 SI 一样，我们使用岭回归估计器，并根据验证集性能调整正则化强度。我们将这一版本的框架称为 SECRETS-VT。

接下来，我们通过将 SECRETS 的假设检验算法替换为标准方法，对
Run Data Driven Hypothesis Test 进行消融实验。我们开展这项消融实验，以评估

在估计的 ITE 之间存在依赖关系的情况下，采用合适的假设检验策略是否必要。首先，我们尝试对正态分布 [17], 的均值进行单样本 t 检验，并将其称为 SECRETS-T；该检验假设底层样本（即 ITE 估计值）是来自高斯分布的独立同分布样本。随后，为评估样本的非高斯性是否会造成问题，我们尝试自助法假设检验，并将其称为 SECRETS-B；该检验仍假设样本相互独立，但放宽了正态性假设，仅要求在零假设下，样本来自参数相同的人群 [18]。我们具体采用偏差校正加速版本 [21] ，因为它能更好地估计感兴趣参数 [18]。随后，为评估样本间的数据依赖是否会造成问题，我们尝试一种 SECRETS-T 的变体，称为 SECRETS-T-P：先对不同试验中的样本进行置换以消除依赖关系，再执行单样本 t 检验。在与这些标准检验策略进行比较后，我们将提出的假设检验策略与“预言机”进行比较，以评估其表现，并将该预言机称为 SECRETS-O。SECRETS 按每个试验构造零分布（通过零假设抽样子程序），而 SECRETS-O 则使用零假设设定下所有试验中的 ITE 估计值来构造零分布，从而确定能够达到 α_{target}显著性水平的正确临界值。

临床随机对照试验数据集

我们在 III 期随机对照试验中评估该框架，因为这类试验通常需要数百至数千名受试者才能确定处理的效果。为了评估我们方法的一般适用性，我们重点考察临床随机对照试验中通常采用的双臂平行组优效性试验设计 [1] 。我们从美国国家神经疾病和中风研究所（NINDS） [27], 获取了三个此类随机对照试验的数据集，即 CHAMP（NCT01581281） [22, 23], 、ICARE（NCT00871715） [24,25], 和 MGTX（NCT00294658） [26], 。这些试验详见附录B。

实现细节

本节将说明用于开展实验的评估参数和计算资源。算法参数详见附录C。

在评估中，我们将 α_{target} 设为 5%；对于样本量要求，我们还将$1-\beta_{target}$ 设为 80%，因为临床随机对照试验中通常采用这一统计工作点 [1]。我们将用于测量检验效能 $1-\beta$ 和显著性水平 α 的试验次数 L 设为 1000，因为根据经验，这足以使各种方法在合理的计算时间内收敛至接近 α_{target} 和 $1-\beta_{target}$ 的水平。对于零假设设定，我们设定 $\mu = \mu_0 = 0$；对于备择设定，我们设定 $\mu = \mu_{1,\ target}$，其中 $\mu_{1,\ target}$ 由各数据集测得的平均处理效应给出，即 CHAMP 为 -3.17、ICARE 为 -3.00、MGTX 为 -2.70。

为运行实验，我们使用了 28–32 个中央处理器核心，每个核心配备 4GB 内存，并采用了主频为 2.4 GHz 的 Intel Broadwell 处理器和主频为 2.6 GHz 的 Intel Skylake 处理器。我们使用 Python 和标准数值计算软件包实现了该框架并开展了实验。

5. 实验结果

本节介绍实验结果。首先，我们将 SECRETS 与基线方法 标准方法进行比较。随后，为了了解 SECRETS 各个组件的重要性，我们评估了消融研究的结果。

SECRETS 与基线方法

在所有数据集中，与标准方法相比，SECRETS 能够获得更高的检验效能和相当的显著性水平，而检验效能的提升意味着所需受试者数量的减少，以便完成。

表1：标准方法 与 SECRETS 的比较，以及 SECRETS 与 SECRETS-VT 在 CHAMP数据集上的比较

n_a	方法	$1-\beta$ (%)	α (%)	$\hat{1-\beta}$ (%)	$\lvert\hat{\mu}_1-\hat{\mu}_0\rvert$	$\hat{\mu}_1-\mu_{1,target}$	$\hat{\mu}_0-\mu_0$	$\hat{\sigma}$ (标准差)	偏移 $term$
550	标准方法	53	5.0	52	3.27	-0.06	0.03	1.63 (0.01)	2.01
	SECRETS	78	5.1	73	1.85	1.31	-0.01	0.72 (0.04)	2.58
	SECRETS-VT	58	5.4	57	3.08	0.07	-0.02	1.43 (0.06)	2.15
1130	标准方法	81	3.8	84	3.20	-0.06	-0.03	1.09 (0.04)	2.93
	SECRETS	97	6.3	95	2.03	1.15	0.01	0.56 (0.04)	3.62
	SECRETS-VT	91	5.8	90	3.10	0.05	-0.02	0.95 (0.03)	3.25

表2：标准方法与 SECRETS，以及 SECRETS 与 SECRETS-VT 在 ICARE数据集上的比较

n_a	方法	$1-\beta$ (%)	α (%)	$\hat{1-\beta}$ (%)	$\lvert\hat{\mu}_1-\hat{\mu}_0\rvert$	$\hat{\mu}_1-\mu_{1,\ target}$	$\hat{\mu}_0-\mu_0$	$\hat{\sigma}$ (标准差)	偏移 $term$
1250	标准方法	27	4.1	30	3.02	-0.01	0.01	2.12 (0.03)	1.43
	SECRETS	81	5.5	81	3.37	-0.33	0.04	1.19（0.03）	2.83
	SECRETS-VT	73	6.5	72	4.96	-1.93	0.04	1.94 (0.02)	2.56
5207	标准方法	81	5.8	81	2.98	-0.04	-0.06	1.05 (0.01)	2.83
	SECRETS	100	5.7	100	3.74	-0.77	-0.02	0.67 (0.07)	5.56
	SECRETS-VT	100	6.2	100	4.87	-1.91	-0.04	0.94 (0.01)	5.16

获得目标检验效能 $1-\beta_{target}$ 和显著性水平 α_{target}。SECRETS 能够始终如一地降低平均处理效应估计值的方差，从而提高检验效能，尽管这会导致在备择设定下平均处理效应估计值的误差更高。例如，在CHAMP数据集（表1）中，当每组样本量（n_a）为 550 时，标准方法 的检验效能（$1-\beta$）为 53%，显著性水平（α）为 5.0%；而 SECRETS 的检验效能为 78%，显著性水平为 5.1%。在给定每组样本量的情况下提高检验效能后，SECRETS 能够在每组样本量为 550 时，接近目标统计工作点：检验效能为 80%（$1-\beta_{target}$），显著性水平为 5%（α_{target}）；而标准方法 则需要每组样本量达到 1130 才能做到。因此，SECRETS 将所需样本数量减少了近 51%，即每组减少 580 名受试者。

为评估 SECRETS 下检验效能提升的影响因素，我们使用功效估计模型（式（1））分析平均处理效应估计值的分布，因为该模型能够很好地估计实测检验效能。例如，在每组样本量为 550 时，该模型估计标准方法和 SECRETS 的检验效能（$1-\beta$）分别为 52% 和 73%，而实测检验效能（$1-\beta$）分别为 53% 和 78%。首先，我们注意到，标准方法下的均值间距更大，即均值之间的距离（$\lvert\hat{\mu}_1-\hat{\mu}_0\rvert$）在标准方法和 SECRETS 下分别为 3.27 和 1.85。与 SECRETS 相比，标准方法的均值区分效果更好，这是因为其在备择设定下的平均处理效应估计值误差较低（$\hat{\mu}_1-\mu_{1,\ target}$），而两者在零假设设定下的平均处理效应估计值误差都较低且相近（$\hat{\mu}_0-\mu_0$）。尽管标准方法下均值区分度更高，但 SECRETS 下的分布方差较低，即标准差平均值（$\hat{\sigma}$）在标准方法和 SECRETS 下分别为 1.63 和 0.72。SECRETS 下较低的方差抵消了标准方法在均值区分度上的优势，使 SECRETS 的位移项更高，为 2.58，而标准方法下为 2.01，进而带来更高的检验效能。每组样本量为 1130 时，以及在 ICARE（表2）和 MGTX 数据集（表3）上，也呈现类似趋势。

在某些数据集上，标准方法 的显著性水平似乎低于目标值5%，而 SE- CRETS 的显著性水平似乎略高于该值（例如表1中 n_a 为1130的 CHAMP），这种性能差距或许可以解释为何 SECRETS 的检验效能高于标准方法。然而，显著性水平上的这一差距可归因于实验局限；预计在消除这些局限后，该差距将会消失，同时保留 SECRETS 带来的检验效能提升。具体而言，

表3：标准方法 与 SECRETS，以及 SECRETS 与 SECRETS-VT 在 MGTX数据集上的比较

n_a	方法	$1-\beta$ (%)	α (%)	$\hat{1-\beta}$ (%)	$\lvert\hat{\mu}_1-\hat{\mu}_0\rvert$	$\hat{\mu}_1-\mu_{1,target}$	$\hat{\mu}_0-\mu_0$	$\hat{\sigma}$ (标准差)	偏移 $term$
30	标准方法	70	5.1	63	2.69	-0.09	-0.10	1.17（0.06）	2.30
	SECRETS	79	5.7	69	1.33	1.33	-0.04	0.54（0.03）	2.45
	SECRETS-VT	71	5.1	77	2.49	0.17	-0.04	0.92（0.01）	2.70
40	标准方法	81	4.8	75	2.67	-0.06	-0.09	1.01 (0.02)	2.63
	SECRETS	87	5.2	80	1.32	1.34	-0.04	0.47 (0.01)	2.80
	SECRETS-VT	85	5.5	87	2.45	0.20	-0.04	0.79 (0.02)	3.09

增加试验次数 L 可使标准方法 收敛至5%的显著性水平，同时保持相当的检验效能（见附录D、表1）。此外，增加 T 或 SECRETS 下在零分布中生成的样本数量，也能使其收敛至5%的显著性水平，同时保持相当的检验效能。例如，在n_a 为1130的 CHAMP数据集上，将 T 从100个样本增加到500个样本，使 α 从6.3%降至5.8%，同时保持检验效能不变；在 n_a 为40的 MGTX数据集上，将 T 从100个样本增加到1K个样本，使 α从20.7%降至5.2%，而检验效能从89.7%降至87.4%。因此，由于标准方法的显著性水平实际上接近5%（基于更多试验），且 SECRETS 的显著性水平会在保持检验效能的同时收敛至5%（通过增加 T），预计在这些情况下 SECRETS 将优于标准方法，功效估计模型也作出了同样的预测（例如表1中 n_a 为1130的 CHAMP）。

SECRETS 的消融研究

在证明 SECRETS 能够提高检验效能，从而降低采样复杂度后，我们评估了该框架中各项设计选择的重要性。首先，通过比较 SECRETS 与 SECRETS-VT 的性能，我们表明，使用 SI 进行 ITE 估计优于使用 VT 进行 ITE 估计。随后，通过与其他标准检验策略进行比较，我们表明，数据驱动的假设检验策略对 SECRETS 的性能至关重要。我们还表明，我们的检验策略表现良好，其性能逐渐接近采用预言机版本策略时所能达到的水平。

Synthetic Intervention vs. Virtual Twins

与 SECRETS 相比，SECRETS-VT 的平均处理效应估计值方差更高，因此其检验效能较低，但总体而言显著性水平相近。例如，在 CHAMP数据集（表1）中，每组样本量为 550 时，SECRETS 和 SECRETS-VT 的显著性水平均接近 5%，但检验效能分别为 78% 和 58%。应用功效方程模型可见，尽管 SECRETS-VT 对均值的区分效果更好，却会增大平均处理效应估计值的方差，从而使其位移项 低于 SECRETS。ICARE数据集（表2）也呈现类似趋势。鉴于 MGTX数据集的样本量较小，该模型并不适用；但 SECRETS-VT 的方差仍高于 SECRETS，这可能解释了其检验效能较低的原因（表3）。

Data-driven Hypothesis Testing vs. Alternative Standard Hypothesis Testing Strategies

接下来，我们将展示：为了在最大化检验效能的同时达到所需的显著性水平 α_{target} （设为 5%），有必要采用能够考虑样本（即 ITE估计值）之间依赖关系的数据驱动假设检验。为此，我们比较了 SECRETS 下的显著性水平

表4：不同假设检验策略下的显著性水平， α (%) s

数据集	n_a	SECRETS-T	SECRETS-B	SECRETS-T-P	SECRETS	SECRETS-O
CHAMP	550	2.2	2.5	6.8	5.1	5.0
	1130	5.3	5.6	5.4	6.3	5.0
ICARE	1250	0.0	0.0	4.3	5.5	5.0
	5207	3.5	3.5	5.8	5.7	5.0
MGTX	30	0.0	0.0	5.3	5.7	5.1
	40	0.0	0.0	5.3	5.2	5.1
平均值	-	1.8	1.9	5.5	5.6	5.0

与标准假设检验策略（即 SECRETS-T、SECRETS-B 和 SECRETS-T-P）下的显著性水平，结果见表 4。SECRETS-T 使用单样本t检验，在各数据集和每组样本量（n_a）下的平均值 α 为 1.8%，这表明该检验的假设——观测值独立同分布于高斯分布——并不成立。首先，我们通过将单样本t检验替换为自助法假设检验，即 SECRETS-B，来检验高斯性是否遭到违反。SECRETS-B 获得了略高的平均值 α （1.9%），说明问题并不在于非高斯性。为证明样本之间存在依赖关系，我们运行了单样本t检验的一个版本，即 SECRETS-T-P；在该版本中，估计的 ITE估计值会在各试验之间进行随机打乱，以消除单次试验内部存在的样本依赖性。SECRETS-T-P 获得了 5.5% 的平均值 α ，显著更接近 α_{target}；这一结果证实，在 SI 下产生的样本依赖性正是标准假设检验所违反的条件。为解决样本依赖性问题，SECRETS 实施了一种利用数据构建零分布的假设检验，该零分布能够捕捉数据中的依赖关系，从而获得 5.6% 的平均值 α 。增加 T 预计有助于通过拟合更准确的零分布，使 SECRETS 的显著性水平趋近于 5%。例如，在每组样本量 n_a 为 1130 的 CHAMP数据集上，将 T 从 100 个样本增加到 500 个样本后， α 从 6.3% 降至 5.8%；在每组样本量 n_a 为 40 的 MGTX数据集上，将 T 从 100 个样本增加到 1K 个样本后， α 从 20.7% 降至 5.2%。此外，SECRETS 的预言机版本，即 SECRETS-O，利用所有试验中的 ITE估计值构建零分布，在所有数据集和每组样本量下均获得了接近 5% 的显著性水平。这表明，SECRETS 可以从零假设下更多样化的 ITE估计值样本中获益，而增加 T可能会获得这些样本。

6. 讨论与未来工作

鉴于实验结果，我们首先讨论其影响，然后提出未来工作的方向。

SECRETS 为什么有效?

我们的实验表明，与基线方法相比，在给定样本量下，SECRETS 能够通过持续降低平均处理效应估计值的方差，在维持显著性水平相当的同时提高检验效能。它通过模拟交叉设计来实现这一点；具体而言，它使用 SI 估计所有个体在对照组和治疗组中的 ITE，并对这些 ITE估计值取平均。我们针对估计量开展的消融研究证实，与标准回归（即 VT 算法）相比，SI 更能降低分布的方差。SI 的优势可能源于它对问题的非参数化表述：它能够学习如何对供体个体的响应进行加权，以预测目标单元的响应；而 VT 算法则仅依据基线测量，使用参数模型预测响应。虽然 SI 对降低平均处理效应估计值的方差至关重要，但 SECRETS 还需要采用适用于 ITE估计值的假设检验算法，以达到所需的显著性水平并最大化

检验效能。由于 SI 使用目标治疗组中的所有个体来预测目标单元的反事实，因此，同一治疗组中目标单元的 ITE估计值会相互依赖，正如我们的假设检验消融研究所示，这使得标准假设检验不再适用。为克服标准假设检验中样本独立性遭到破坏的问题，SECRETS 使用一种新颖的数据驱动假设检验算法，该算法构造一个能够刻画 ITE估计值之间依赖关系的零分布，且不对其底层分布作任何参数假设。因此，有效的反事实估计器与适当的假设检验方案相结合，正是 SECRETS 的“动力”所在。

未来工作

为提高 SECRETS 的效能和实用性，我们提出一些进一步研究的方向。首先，可以改进反事实估计算法，以降低平均处理效应估计值的误差，尤其是在备择设定下，因为研究也可能关注平均处理效应估计值。一个可行的方法是基于所有测得的基线协变量训练估计量，因为临床随机对照试验通常会收集大量有序和数值测量值，涉及人口统计学特征、病史和临床评估等类别（例如，我们考察的随机对照试验包含数百至数千个此类协变量）。虽然 SI 只需将额外协变量补充到干预前数据中，就能纳入多个协变量，但这种策略需要加以调整，才能适用于不同类型的协变量，即数值和有序协变量，因为不同协变量类型对应的回归设定各不相同 [28]。除了提升估计量的性能外，还可以为 SECRETS 增设一个模块，估算达到预期显著性水平和检验效能所需的样本量，从而提高其在研究规划中的实用性。

7. 结论

总之，我们开发了 SECRETS 这一全新框架。该框架首次在不依赖任何外部数据的情况下提高了临床随机对照试验的检验效能，从而降低了抽样复杂度，这与以往方法不同。SECRETS 通过模拟交叉设计实现这一目标：它针对对照组和治疗组中的每位患者估计个体处理效应，以降低平均处理效应估计值的方差；而以往方法仅对治疗组进行此类估计。为估计两组患者的个体处理效应，SECRETS 使用了先进的反事实估计算法 SI。随后，为利用平均处理效应估计值方差的降低来提高检验效能，SECRETS 实施了一种全新的数据驱动假设检验策略，适用于估计的个体处理效应，因为在标准假设检验方案下，这些估计值的性质不符合独立性假设。在三项真实世界 III 期临床随机对照试验（即 CHAMP、ICARE 和 MGTX 研究）中进行评估后发现，与基线方法相比，SECRETS 将检验效能提高了 6%–54%，平均提升 21.5%（标准差为 15.8%），同时保持了相当的显著性水平。此外，检验效能的提升减少了达到通常期望的统计工作点（80% 检验效能和 5% 显著性水平）所需的受试者数量：在 CHAMP 数据集中，每组减少 51%（即 580 名受试者）；在 ICARE 数据集中，每组减少 76%（即 3957 名受试者）；在 MGTX 数据集中，每组减少 25%（即 10 名受试者）。目前，SECRETS 可用于提升已完成临床随机对照试验的检验效能。未来，我们将着力改进反事实估计算法，以减少平均处理效应估计值中的误差，并开发一种基于 SECRETS 的样本量估计方法，作为研究规划的辅助工具。

致谢

本研究由美国国家科学基金会资助（资助编号：CNS-1907381），并使用了普林斯顿研究计算提供的资源。

本研究基于从美国国家神经疾病和中风研究所临床研究存档数据集网站获取的数据。CHAMP数据集来自儿童及青少年偏头痛预防研究，该研究由主要研究者 Powers、Hershey 和 Coffey 医生负责，资助编号#1U01NS076788-01。ICARE数据集来自卒中后上肢康复研究（ICARE），该研究由主要研究者 Winstein、Dromerick 和 Wolf 医生负责，资助编号#U01NS056256。MGTX数据集来自非胸腺瘤性重症肌无力患者接受泼尼松治疗的胸腺切除术试验，该研究由主要研究者 Cutter、Wolfe 和 Kaminski 医生负责，资助编号#1U01NS042685-01A2。

作者感谢 Wolfe 博士在随机对照试验和 MGTX 研究方面给予的指导。

参考文献

[1] L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin 和 C. B. Granger，《临床试验基础》。Springer，2015 年。[2] L. Yao、Z. Chu、S. Li、Y. Li、J. Gao 和 A. Zhang，“因果推断综述”，《*ACM* 数据知识发现汇刊》，第 15 卷，第 5 期，第 1–46 页，2021 年。[3] K. Stanley，“随机对照试验的设计”，《循环》，第 115 卷，第 9 期，第 1164–1169 页，2007 年。[4]“美国食品药品监督管理局药物审批流程。” [在线]。获取地址：https://www.fda.gov/media/82381/ download[5] V. Prasad，“可靠、便宜、快速且数量少：评估医疗实践的最佳研究是什么？随机对照试验还是合成对照组？”《欧洲临床研究杂志》，第 51 卷，第 8 期，2021 年。[6] S. S. Senn，《临床研究中的交叉试验》。约翰·威立父子出版公司，2002 年。[7] J. W. Blackston、A. G. Chapple、J. M. McGree、S. McDonald 和 J. Nikles，“通过模拟研究比较汇总单病例试验与平行和交叉随机对照试验”，《医疗保健》，第 7 卷，第 4 期，第 137 页，2019 年。[8] K. Thorlund、L. Dron、J. J. Park 和 E. J. Mills，“临床试验中的合成对照和外部对照：研究人员入门指南”，《临床流行病学》，第 457–467 页，2020 年。[9] I. Bica、A. M. Alaa、C. Lambert 和 M. Van Der Schaar，“从真实世界患者数据到利用机器学习估计个体化治疗效应：应对潜在挑战的当前和未来方法”，《临床药理学与治疗学》，第 109 卷，第 1 期，第 87–100 页，2021 年。[10] E. Højbjerre-Frandsen、M. L. Jeppesen 和 R. K. Jensen，“利用数字孪生提高随机临床试验的检验效能”，硕士论文，奥尔堡大学，2022 年。[11] J. Lim、R. Walley、J. Yuan、J. Liu、A. Dabral、N. Best 等，“在创新性确证性临床试验中使用历史受试者级数据以减轻患者负担：方法与机遇综述”，《治疗创新与监管科学》，第 52 卷，第 546–559 页，2018 年。[12] Z. Qian、Y. Zhang、I. Bica、A. Wood 和 M. van der Schaar，“Synctwin：利用纵向结局估计治疗效应”，载于第 *34* 届《神经信息处理系统进展》论文集，2021 年。[13] A. Agarwal、D. Shah 和 D. Shen，“利用合成干预进行合成A/B测试”，2023 年，arXiv:2006.07691v5。

[14] J. C. Foster、J. M. Taylor 和 S. J. Ruberg，“基于随机临床试验数据的亚组识别”，《医学统计学》，第 30 卷，第 24 期，第 2867–2880 页，2011 年。[15] J. M. Wolf、J. S. Koopmeiners 和 D. M. Vock，“一种在控制I类错误率的同时检测随机临床试验中异质性治疗效应的置换程序”，《临床试验》，第 19 卷，第 5 期，第 512–521 页，2022 年。[16] M. Amjad、D. Shah 和 D. Shen，“稳健合成控制”，《机器学习研究杂志》，第 19 卷，第 1 期，第 802–852 页，2018 年。[17] B. Rosner，《生物统计学基础》。圣智学习，2015 年。[18] P. Good，《假设的置换检验、参数检验与自助法检验》。Springer Science +Business Media, Inc.，2005 年。[19] M. R. Dale 和 M.-J. Fortin，“空间自相关与生态学中的统计检验”，《生态科学》，第 9 卷，第 2 期，第 162–167 页，2002 年。[20] M.-J. Fortin 和 G. M. Jacquez，“随机化检验与空间自相关数据”，《美国生态学会通报》，第 81 卷，第 3 期，第 201–205 页，2000 年。[21] B. Efron，“更优的自助法置信区间”，《美国统计学会杂志》，第 82 卷，第 397 期，第 171–185 页，1987 年。[22] S. W. Powers、C. S. Coffey、L. A. Chamberlin、D. J. Ecklund、E. A. Klingner、J. W. Yankey等，“阿米替林、托吡酯和安慰剂治疗儿童偏头痛的试验”，《新英格兰医学杂志》，第 376 卷，第 2 期，第 115–124 页，2017 年。[23] “儿童与青少年偏头痛预防研究（CHAMP）。” [在线]。获取地址：https://clinicaltrials.gov/ct2/show/results/NCT01581281[24] C. J. Winstein、S. L. Wolf、A. W. Dromerick、C. J. Lane、M. A. Nelsen、R. Lewthwaite 等，“任务导向康复计划对运动性卒中后上肢功能恢复的影响：ICARE随机临床试验”，《美国医学会杂志》，第 315 卷，第 6 期，第 571–581 页，2016 年。[25] “卒中后上肢康复研究ICARE。” [在线]。获取地址：https://clinicaltrials.gov/ ct2/show/results/NCT00871715[26] G. I. Wolfe、H. J. Kaminski、I. B. Aban、G. Minisman、H.-C. Kuo、A. Marx 等，“胸腺切除术治疗重症肌无力的随机试验”，《新英格兰医学杂志》，第 375 卷，第 6 期，第 511–522 页，2016 年。[27]“已归档临床研究数据集。” [在线]。获取地址：https://www.ninds.nih.gov/
c
u
r
re
nt-research/research-funded-ninds/clinical-research/archived-clinical-research-datasets
[28] F. E. Harrell Jr.，《回归建模策略：线性模型、逻辑回归和生存分析的应用》。Springer，2001 年。

附录 A。SECRETS 伪代码

算法 1 SECRETS

输入:
$X_{ctrl} \in \mathbb{R}^{n_a \times n_t}$：对照组数据，其中 n_a 为每组样本量， n_t 为研究持续时间，其中 $t = 1$ 为干预前时间点， $t > 1$ 为干预后时期

$X_{treat} \in \mathbb{R}^{n_a \times n_t}$：治疗组数据，其中 n_a 为每组样本量， n_t 为研究持续时间，其中 $t = 1$ 为干预前时间点， $t > 1$ 为干预后时期

周期
SI_调优参数：包含以下 SI 调优参数的字典
$r_{train,val}$：用于调优 SI 超参数的训练集与验证集大小之比
$get_outcome$： $\mathbb{R}^{n_t} \to \mathbb{R}$：根据患者在研究时长内接受某种干预后的响应轨迹计算感兴趣结果的函数 n_t
T：从零分布生成的样本数
检验参数：用于调整测试的字典，包含以下参数。
临界值
α_{target}：目标显著性水平
t_{lower}：临界值搜索下界； ≥ 0 （由于采用双侧检验）
t_{upper}：临界值搜索上限； $\geq t_{lower}$
$t_{limit,\ exp}$：临界值搜索限制扩展术语
n_s：待搜索的候选临界值数量
$\delta_{\alpha_{target}}$： 显著性水平误差容限

输出:
d：检验结果，1 表示拒绝零假设，0 表示不拒绝零假设1: $Y_{ctrl} =$ ESTIMATE ITES(X_{ctrl}, X_{treat},SI 调优参数 $get\ outcome$)2: $Y_{treat} =$ ESTIMATE ITES(X_{treat}, X_{ctrl},SI 调优参数 $get\ outcome$)3: $Y = concatenate(Y_{ctrl}, Y_{treat})$4: $d =$ RUN HYPOTHESIS TEST(X_{ctrl}, Y, SI 调优参数 $get\ outcome, T$,检验参数)

5: **返回** d

算法 2 TUNE SI HYPERPARAMS

输入:
$X \in \mathbb{R}^{n_a \times n_t}$：来自接受单一干预的一组受试者的数据，其中 n_a 是该组的受试者人数，其中 n_t 是研究时长， $t = 1$ 是
干预前时间点， $t > 1$ 是干预后时期
params：包含以下 SI 调优参数的字典
$r_{train,val}$：训练集与验证集大小之比，用于调优 SI 的超参数

输出:
$\lambda_{ridge,\ best}$：调优后的岭正则化强度
$\lambda_{svt,best}$：奇异值阈值化的调优后截断阈值

1: X_{train}， $X_{val} = split\ train_val(X,$ 参数$.r_{train,\ val})$
2: $size,\ n_t = shape\ (X_{val})$
3: $best_score = -\infty$
4: $\lambda_{ridge,\ best} = \mathrm{NIL}$
5: $\lambda_{svt,best} = \mathrm{NIL}$
6: **对于** $\lambda_{ridge} \in logspace(-3,\ 3,\ 7)$ **执行**
7: **针对** $\lambda_{svt} \in linspace(0。1,\ 1,\ 10)$ **执行**
8: $X_{val,pred} = \mathbf{0}^{size \times n_t}$
9: **针对** $i = 1, size$ **执行**
10: $X_{val,pred}[i]$=SI(λ_{ridge},λ_{svt},X_{train},$X_{val}[i]$)
11: **结束循环**
12: $score = get_rsquared(X_{val}, X_{val,pred})$
13: **如果** $score > best_score$，**则执行**
14: $best_score = score$
15: $\lambda_{ridge,best} = \lambda_{ridge}$
16: $\lambda_{svt,best} = \lambda_{svt}$
17: **结束 if**
18: **结束循环**
19: **结束于**
20: **返回** $\lambda_{ridge,\ best}$， $\lambda_{svt,\ best}$

TUNE SI HYPERPARAMS (算法2)首先初始化最佳超参数集合（第3–5行），然后在 λ_{ridge} 和 λ_{svt}的候选值范围内进行广泛搜索，并跟踪每种超参数配置下的验证性能（第6–19行）。验证性能是指真实验证数据与相应SI派生预测之间的 R^2 评分。为获取每个个体的SI派生预测，
TUNE SI HYPERPARAMS 使用给定的候选超参数运行 SI (算法3)，并将 X_{train} 用作供体数据、$X_{val}[i]$ 用作目标单元数据（第9–11行）。在估计验证单元的轨迹后， TUNE SI HYPERPARAMS 计算验证性能；如果验证性能有所提升，则更新最佳超参数配置（第12–17行）。最后，返回搜索范围内的最佳超参数配置。

算法 3 SI

输入:
λ_{ridge}: 岭正则化强度
λ_{svt}: 奇异值阈值化的截断阈值
$X_{donor} \in \mathbb{R}^{n_d \times n_t}$: 供体数据，其中 n_d 是供体受试者的数量， n_t 是研究持续时间，其中 $t = 1$ 是干预前时间点， $t > 1$ 是干预后时期
$X_{unit} \in \mathbb{R}^{n_t}$: 目标单元数据，其中 n_t 为研究时长，其中 $t = 1$ 为干预前时间点， $t > 1$ 为干预后时期

输出:
$X_{unit,\ c} \in \mathbb{R}^{n_t}$: 目标单元的反事实数据（在分配给供体组的干预下），贯穿整个研究时长

1: $n_d,\ n_t = shape(X_{donor})$
2: $X_{unit} = reshape(X_{unit}, (1, n_t))$
3: $U,\ s,\ V^T = \text{SVD}\ (X_{donor})$
4: $S = \{i :\ s[i] \geq \lambda_{svt}\}$
5: $X_{donor,\ trunc} = \sum_{i \in S} s[i]U[i]V[i]^T$
6: $w =$ 参数最小值$_w \|X_{unit}[:\ 1] - w^T X_{donor,\ trunc}[:\ 1]\|_2^2 + \lambda_{ridge}\|w\|_2^2$
7: $X_{unit} = \mathbf{0}^{n_t}$
8: $X_{unit,c}[1] = squeeze(X_{unit})[1]$
9: $X_{unit,\ c}[2:] = squeeze(w^T X_{donor,\ trunc}[:, 2:])$
10: **返回** $X_{unit\ \ ,c}$

SI (算法3)通过对供体数据进行奇异值阈值化去噪（第3–5行），在干预前（即基线）数据上使用岭回归学习供体权重（第6行），并使用学习到的供体权重，根据供体的干预后数据估计目标单元的干预后轨迹（第9行），从而估计目标单元的反事实干预后轨迹。由于 SI 用于预测目标单元的干预后轨迹，因此将该单元的“反事实”干预前数据初始化为其原始干预前数据（第8行）。

算法 4 TUNE CRITICAL VALUE_

输入:
$s_{null} \in \mathbb{R}^T$：来自检验统计量零分布的 T 个样本。params：包含以下用于调整检验的参数的字典 临界值α_{target}：目标显著性水平t_{lower}：临界值搜索下界； ≥ 0 （因为是双侧检验）t_{upper}：临界值搜索上界； $\geq t_{lower}t_{limit,\ exp}$：临界值搜索限制扩展术语$n_s$：要搜索的候选临界值数量$\delta_{\alpha_{target}}$：显著性水平误差容限

输出:
d：检验结果，1 表示拒绝零假设，0 表示不拒绝零假设 1：过程 GET ALPHA (s_{null}, t) 2：$\alpha = 0$3：对于 $i = 1, length(s_{null})$ 执行4 $r = \left|s_{null}[i]\right| > t$：：5 $\alpha = \alpha + r$ 6：结束循环 7：返回 $\alpha/length(s_{null})$ 8：结束过程 9： $t_{candidates} = linspace$ (params.t_{lower}, params.t_{upper}, params.n_s) 10：$\hat{}\alpha\ by\ t = []$11：对于 $i = 1$, params.n_s 执行 12：$t = t_{candidates}[i]$13：$\hat{}\alpha$=GET ALPHA(s_{null}, t) 14：如果 $|\hat{}\alpha -$ params.$\alpha_{target}|<$ params.$\delta_{\alpha_{target}}$ 则 15：返回 t16：结束 if 17：$\hat{}\alpha\ by\ t[i] = \hat{}\alpha$18：结束循环 19： $m =$ params.$n_s/2$20： $\alpha_m = \hat{}\alpha\ by\ t[m]$21：如果 params.$\alpha_{target} < \alpha_m$ 则 22：params.$t_{lower} = t_{candidates}[m + 1]$23：如果 params.$\alpha_{target} < \hat{}\alpha\ by\ t[-1]$ 则 24：params.$t_{upper} =$ params.$t_{upper} +$ params.$t_{limit,exp}$25：结束 if 26：否则 27：params.$t_{upper} = t_{candidates}[m + 1]$28：如果 params.$\alpha_{target} > \hat{}\alpha\ by\ t[0]$ 则 29：params.$t_{lower} = max(0,$ params.$t_{lower} -$ params.$t_{limit,exp}$) 30：结束 if 31：结束 if 32：返回 TUNE CRITICAL VALUE(s_{null}, $params$)

TUNE CRITICAL VALUE （算法 4）通过搜索临界值，找出能达到 params.α_{target} 所指定显著性水平的临界值。首先，它遍历由 params.n_s、params.t_{lower} 和 params.t_{upper} 定义的一系列候选临界值 $t_{candidates}$，并检查是否有某个值能达到所需的显著性水平 params.α_{target} （第 9–18 行）。对于每个 t，它通过调用 GET ALPHA （第 13 行）估计相应的显著性水平 $\hat{\alpha}$；该调用通过使用候选临界值 t 进行双侧检验来计算 $\hat{\alpha}$。GET ALPHA （第 1–8 行）对零分布 s_{null} 中的每个样本（第 4 行）执行检验，并计算检验返回拒绝结果的样本比例（第 7 行）。如果该临界值所得的 $\hat{\alpha}$ 接近 params.α_{target} （即符合误差容限 params.$\delta_{\alpha_{target}}$），TUNE CRITICAL VALUE 就返回该临界值（第 14–16 行）；否则，它会保存相应的显著性水平并继续遍历（第 17 行）。

如果没有任何候选值能产生接近 params.α_{target} 的显著性水平，TUNE CRITICAL VALUE 会采用类似二分搜索的方式更新其搜索范围。首先，它会确定中间候选临界值 α_m 对应的显著性水平（第 19–20 行）。如果 params.α_{target} 小于 α_m，TUNE CRITICAL VALUE 会将搜索范围更新为原始临界值搜索空间的右半部分，因为增大临界值会降低显著性水平（第 21–22 行）。此外，如果 params.α_{target} 小于与 params.t_{upper} 相关联的 $\hat{\alpha}$，则 params.t_{upper} 需要充分增大（即增加 params.$t_{limit,\ exp}$），以确保更新后的搜索范围包含一个能产生 α_{target} 的临界值（第 23–24 行）。类似地，如果 params.α_{target} 大于 α_m，TUNE CRITICAL VALUE 会将搜索范围更新为原始临界值搜索空间的左半部分，因为减小临界值会提高显著性水平（第 26–27 行）。此外，如果 params.α_{target} 大于与 params.t_{lower} 相关联的 $\hat{\alpha}$，则 params.t_{lower} 需要充分减小（即减小 params.$t_{limit,\ exp}$，但由于双侧检验，其下限为 0），以确保更新后的搜索范围包含一个能产生 α_{target} 的临界值（第 28–29 行）。更新搜索范围后，TUNE CRITICAL VALUE 会使用更新后的搜索参数递归调用自身，继续搜索并返回已确定的临界值（第 32 行）。

附录B. 临床随机对照试验数据集

B.1 CHAMP

CHAMP研究 [22] 评估了不同药物能否降低患有偏头痛的儿童和青少年的头痛频率及头痛影响。为评估这一点，该研究实施了一项随机对照试验，设有一个安慰剂组和两个治疗组，分别接受阿米替林和托吡酯。试验结束时，研究测量了各项结果的治疗效应，包括头痛频率的变化、头痛天数以及头痛相关残疾评分，均以基线测量为参照。根据对328名受试者为期24周收集的数据，该研究未发现各项健康结局在组间存在具有临床意义的差异。

在我们的实验中，我们使用儿童偏头痛残疾评估（Ped-MIDAS）评分的变化作为定义平均处理效应的结局指标，并分别将阿米替林和托吡酯作为对照组和治疗组，因为与其他连续指标和试验组配对所定义的平均处理效应相比，相应的平均处理效应具有更高的统计显著性 [22, 23]。为计算这一平均处理效应，该研究分析了在两次访视期间接受监测的一部分受试者，即基线时一次，另一次在 24 周终点附近；这部分受试者共 211 人，其中阿米替林组 107 人，托吡酯组 104 人。采用相同标准后，我们从原始数据集中提取了 204 名受试者，其中阿米替林组 106 人，托吡酯组 98 人。根据提取的数据集，我们计算得出的平均处理效应为 -3.17 个体，与研究中报告的 -4.3 个体大致相当 [23]。

B.2 ICARE

ICARE研究 [24] 评估了一项新的运动训练计划（加速技能习得计划，即ASAP）是否能够比常规护理（UCC）更有效地减轻运动性卒中患者的上肢残疾。为此，该研究实施了一项随机对照试验（RCT），其中包括接受ASAP的治疗组、接受剂量等效常规护理（DEUCC）的对照组，以及接受不限制剂量的常规护理的另一对照组。试验结束时，研究评估了多个结局指标的治疗效应，包括Wolf运动功能测试（WMFT）用时的变化、卒中影响量表（SIS）评分以及手臂肌肉扭矩。基于对304名患者进行为期一年的数据收集，研究未发现部分评分存在具有临床意义的组间差异，即WMFT和SIS手功能评分 [24, 25]。

在我们的实验中，我们使用基于肩屈肌的手臂肌肉扭矩变化作为定义平均处理效应（ATE）的结局指标，并分别将DEUCC和ASAP作为对照组和治疗组，以缩短实验时间，因为对于基线方法而言，以较高检验效能和较低显著性水平检测该ATE只需要相对较小的样本量。在对原始数据集应用该研究的数据处理方案后，我们从基线和一年终点的访视中提取了183名受试者的数据，其中93名属于对照组，90名属于治疗组（每组分析的受试者数量与研究报告中的分析数量相同）。根据提取的数据集，我们计算得出的ATE为-3.00个体，与研究报告的-2.99接近 [25]。

B.3 MGTX

MGTX研究 [26] 探讨了胸腺切除术联合标准泼尼松治疗是否比单独使用泼尼松治疗更有效地治疗重症肌无力。为此，该研究开展了一项随机对照试验：对照组在三年内接受泼尼松治疗，治疗组则接受胸腺切除术和相同的泼尼松治疗；试验结束时，研究根据定量重症肌无力

（QMG）总评分和所需的泼尼松剂量来衡量治疗效应，两者均按整个研究期间的平均值计算。根据对126名受试者三年间收集的数据，研究发现胸腺切除术在临床上显著改善了健康结局；时间加权平均QMG评分平均降低了2.85个个体，时间加权平均泼尼松剂量也平均减少了22毫克。

在我们的实验中，我们使用按时间加权的 QMG 总评分作为定义平均处理效应的结局指标，因为这种指标更容易复现 [26]。该研究通过分析三年时间窗口内监测的人群子集（14 次患者就诊或时间点）来计算平均处理效应，其中治疗组和对照组分别包含 62 名和 56 名受试者。我们遵循该研究的数据处理方案，并从原始数据集中提取了一个数据集，其中治疗组包含 49 名受试者，对照组包含 47 名受试者。我们根据提取的数据集计算得出平均处理效应为 -2.70 个单位，与该研究报告的 -2.85 个单位相近。

附录C. 算法参数

对于每个随机对照试验数据集，我们根据附录 B 中指定的相应结局指标定义结局获取函数，供标准方法和 SECRETS 使用。对于标准方法，我们将目标显著性水平设为 $\alpha_{target} = 5\%$。对于 SECRETS，我们将 SI 调优参数中包含的参数设置如下：$r_{train,\ val}$ 设为 7/3。我们将 T 设为 CHAMP 和 ICARE 的 100，以及 MGTX 的 1000。我们将检验参数中包含的参数设置如下：$\alpha_{target} = 5\%$、$t_{lower} = 3$、$t_{upper} = 5$、$t_{limit,\ exp} = 2$、$n_s = 10$ 和 $\delta_{\alpha_{target}} = 1e\text{-}3$。我们发现，这种参数配置使 SECRETS 达到接近 α_{target} 的显著性水平。有关这些参数的说明，请参见附录 A 中的算法 1。

附录D. 表格

表1：试验次数（L）对标准方法性能的影响

数据集	$L = 1$K		$L = 10$K	
	$1-\beta$ (%)	α (%)	$1-\beta$ (%)	α (%)
CHAMP , $n_a = 1130$	81	3.8	81	4.5
ICARE , $n_a = 1250$	27	4.1	28	4.9
ICARE , $n_a = 5207$	81	5.8	80	5.5