arXiv:2302.03544v1 [stat.ME] 7 Feb 2023

因果可解释随机效应荟萃分析

Justin M. Clark* Kollin W. Rott* James S. Hodges* Jared D. Huling*

摘要

近期研究在因果可解释荟萃分析的发展方面作出了重要贡献。这些方法将从一组随机试验中估计的治疗效应迁移至所关注的目标人群。理想情况下，针对特定人群的估计值更易解释，也与政策制定者和临床医生更为相关。然而，并非源于治疗效应修饰因子分布差异的研究间异质性，可能会增加跨试验综合估计值的难度。这类异质性的存在（包括治疗方式的差异）也使得将迁移估计值解释为目标人群中的一般性效应变得更加复杂。我们提出了一个概念框架和估计程序，试图对这类异质性加以考虑，并制定了旨在捕捉因果估计值中伴随的额外变异的推断技术。该框架还旨在阐明哪些类型的治疗效应适用于可推广性和可迁移性技术。

关键词：因果推断、可推广性、可迁移性、荟萃分析、证据综合、临床试验

*明尼苏达大学　公共卫生学院，生物统计学部

学

1 引言

基于一组随机对照试验的数据，临床医生和政策制定者都面临一个重要问题：这些结果是否适用于目标人群。近期研究通过开发专为荟萃分析设计的因果推断方法，在解决这一问题上取得了重要进展（Dahabreh 等人，2022）。这些方法旨在考虑试验人群与目标人群之间的差异，从而得出对目标人群具有因果解释的效应估计。

此类人群差异的影响，是荟萃分析临床医生早已认识到的研究综合挑战之一：评估不同试验的结果时，应考虑各种类型的研究间异质性。人群特征差异只是研究间异质性的一种形式。本文在现有因果可解释荟萃分析框架的基础上，提出因果量和估计量，以进一步考虑未得到解释的研究间异质性；这种异质性不仅源于试验集合与目标人群之间的协变量差异。

我们的方法所产生的效应估计既适用于具有科学研究价值的目标人群，也能在研究间异质性导致无法合并各试验数据时保持可解释性。我们还认为，这项工作在新兴的因果可解释荟萃分析领域与证据综合中已得到充分发展的随机效应方法之间搭建了一座概念桥梁。这种与传统随机效应荟萃分析的结构相似性，使我们能够采用该方法中的一些重要分析框架，同时保留因果可解释性。

原则上，随机试验的系统综述应对卫生政策和临床实践产生重要影响（Berlin and Golub 2014）。例如，推荐强度分类法（SORT）将荟萃分析列为研究质量的最高等级（Ebell et al. 2004）。然而，近期研究表明，荟萃分析——尤其是涉及个体患者数据（IPD）的荟萃分析——对制定和发布临床指南的组织影响相对有限（Vale et al. 2015）；即使对于那些能够轻易获得相关 IPD 荟萃分析的指南而言，也不例外。

2021年发表的一项类似研究也发现，系统综述在为临床指南提供依据方面应用有限：在该研究中，接受分析的指南里只有34%对现有证据进行了系统综述，以指导临床实践（Lunny 等，2021年）。这些作者为其研究结果提出了许多可能的解释，包括指南制定组缺乏相关认识、因此过度依赖专家意见，以及完成高质量系统综述所需的大量劳动和时间。然而，值得探究的是，系统综述在为临床指南提供依据方面应用不足，是否可能源于荟萃分析本身的实用性

不足。也就是说，对临床证据的综述可能具有系统性和严谨性，并符合国际标准，但仍可能不太适用于特定人群或环境中的决策。

某项试验对临床实践的相关性，很大程度上取决于临床医生所针对的患者人群与试验参与者的契合程度。招募过程复杂且因研究而异，可能导致研究虽具有较高的内部效度，却不太可能适用于与试验参与者不同的个体，进而降低外部效度（Degtiar and Rose 2021）。外部效度问题与卫生系统的决策尤为相关，因为受益人可能在重要方面不同于用于支持医保覆盖决策的随机对照试验中的参与者。例如，2008年的一项研究指出，为美国医疗保险和医疗补助服务中心政策提供依据的一项荟萃分析所纳入的随机对照试验，其参与者与美国医疗保险人群存在显著差异（Dhruva 2008）。这种内部效度与外部效度之间的落差，或许有助于解释为何人们日益关注利用大型电子健康记录（EHR）数据库为临床研究提供信息（Galson and Simon 2016）。

在综合多项研究的证据时，相关性和外部效度问题更加突出。许多挑战都与某项荟萃分析所纳入的随机对照试验之间可能存在的异质性有关。仅考虑一种可能的差异，即各项随机对照试验所涉及的参与者人群存在异质性，就很难想象标准荟萃分析中对这些人群取平均值而得出的治疗效应估计，如何能够适用于某位临床医生或某个卫生系统所关注的人群。此外，治疗方式以及主要结局的评估方法等其他相关异质性来源依然存在（Berlin and Golub 2014）。

对此类研究间异质性进行建模的标准方法是随机效应荟萃分析，在这种方法中，每项研究的效应或结局都被视为从某种分布（通常是正态分布）中随机抽取的结果（Higgins、Thompson 和 Spiegelhalter, 2009）。因此，试验间的异质性被视为试验结局总方差的一部分：其中一个部分源于试验内的抽样变异，另一个部分则反映试验间的系统性（但仍为独立同分布）变异。

然而，随机效应荟萃分析得出的估计值仍未能明确考虑试验参与者与临床医生和政策制定者所关注的目标人群之间的差异。近期因果可解释荟萃分析领域的研究提出了利用多个随机对照试验的数据进行人群特异性推断的方法（Dahabreh 等人（2022）、（2020））。大体而言，这些研究利用目标人群的代表性样本，将在随机对照试验人群中估计的治疗效应迁移至目标人群。

这种方法的一个关键假设是，来自同一目标人群的两个个体，无论参与荟萃分析中的哪项临床试验，其治疗的平均反应都相同

。此外，假设目标人群中的某个个体若参与任意一项试验，其观察到的治疗反应与更广泛人群中预期的治疗反应相似。试验实施中的系统性异质性以及试验参与效应的可能性，都会对这两个假设构成威胁（Dahabreh、Robertson 和 Hernán，2022）。虽然部分异质性可能由协变量差异解释，但其他特定研究地点机制可能与参与者特征无关。

1.1 研究间异质性的示例

为了更具体地界定此类异质性，可以设想这样一种情形：荟萃分析中的每项研究都采用了略有不同的治疗版本。例如，在旨在缓解下背痛的研究中，一项研究的治疗提供者可能采用与另一项研究的提供者略有不同的脊柱手法治疗形式。这些差异可能已在各项研究的方案中预先规定，但也可能仅仅是因为不同的脊椎按摩治疗师实施手法治疗的方式略有不同而产生。此外，疼痛结局的测量方式不同（例如，相对于治疗，评估疼痛量表的确切时间不同）也可能需要扩展潜在结果的记法。

例如，治疗版本方面的这些差异，即使在考虑了研究之间以及试验与目标人群之间治疗效应修饰因子分布的差异之后，仍然会存在。在下背痛的例子中，如果两名协变量数据完全相同的个体分别参与了不同的研究，那么在尝试比较并界定他们各自的潜在结果时，脊柱手法治疗类型的差异仍然存在。这表明，应当在潜在结果本身中标记此类异质性，例如，某个个体在某位治疗提供者而非另一位治疗提供者处接受分配的治疗 a 时的潜在结果。

试验实施方面更广泛的差异也可能导致这种异质性。例如，在 COVID-19大流行扰乱生物医学研究后，一些临床试验选择通过邮寄方式向试验参与者寄送自我给药药物，而不是当面发放（美国食品药品监督管理局，2021年）。在这些情况下，不同试验的药物送达时间可能有显著差异。例如，参与者居住在较偏远农村地区的试验，其送达时间可能比在城市地区开展的试验更长。对于 COVID-19 这类疾病，治疗时机会影响后续结局，因此送达时间上的差异可能导致各项试验的潜在结局出现异质性。

面对面施用治疗的时间差异也可能对研究综合产生重要影响。近期的一篇预印本考察了被动施用的抗体是否会改变与SARS-CoV-2感染相关的健康结局（Stadler等，2022）。研究人员发现，更早施用

此类治疗可能会提高其疗效。设想两名基线协变量完全相同的患者分别参加两项研究，这两项研究考察的是同一种单克隆抗体治疗。如果两项研究中治疗相对于感染的施用时间不同，这项研究表明，这些个体的预期潜在结局可能各不相同，并且具有研究特异性。

试验实施方面的差异可能导致异质性，另一个例子是将提供者评定的评分作为结局。例如，阳性与阴性症状量表（PANSS）常用于评估精神分裂症患者的症状严重程度（Kay、Fiszbein 和 Opler，1987）。对于 PANSS 评估的每种症状，评分者会给出 1 分（无症状）至 7 分（极重）的评分。随后，可以通过比较治疗组的 PANSS 汇总分数来评估组间差异。然而，不同评分者进行这类定性评估的方式可能有所不同。综合采用不同评分者的多个地点的结局时，分析可能需要将这些差异纳入考量。

受这些问题的启发，我们的工作旨在为建立一种因果可解释的荟萃分析奠定基础，使其能够考虑研究间的异质性，而不仅限于治疗效应修饰因子分布差异所导致的异质性。我们的方法尝试将传统随机效应荟萃分析的技术与直觉，同 Dahabreh 等人（2022）提出的基于因果推断的荟萃分析框架结合起来。具体而言，这涉及提出两个因果问题。首先，如果目标人群中的成员参与了纳入相关随机对照试验集合的某项特定研究，我们平均会观察到怎样的结局？此外，如果我们开展一项新的随机对照试验，且其参与者直接来自目标人群，我们会观察到怎样的结果？后一个问题是本研究主要关注的科学问题，并直接源自我们的新框架。通过回答此类问题，我们希望扩展卫生政策制定者和临床医生在通过荟萃分析评估治疗效果时可用的工具。

2 Es用于捕捉研究间异质性的 estimand y

因果可解释荟萃分析通过可迁移性和可推广性的视角，分析一组研究中观察到的结局和治疗效应。我们假设存在某个感兴趣的目标人群，并探讨这组研究能够告诉我们干预措施在该目标人群中的效应。假设我们拥有来自一组临床试验的数据，这些试验由集合 $\mathcal{S} = \{1, ..., m\}$进行索引。一个感兴趣的因果问题是：如果目标人群中的某位成员参加了研究 $s \in \mathcal{S}$，我们可以预期其获得怎样的治疗效应？正如我们稍后将更精确地阐述的那样，我们可以大致将这种平均治疗效应 τ_s 描述为总体均值 τ 与

研究特异性偏差 δ_s s:

$$\tau_s = \tau + \delta_s. \tag{1}$$

在这种表述中，总体均值 τ 给出了目标人群中的治疗效应在这组研究上的平均值，而 δ_s 反映了潜在治疗效应的研究间异质性。

我们的最终目标是精确刻画、估计并对 τ_0 进行推断；如果在目标人群中开展一项新试验，这就是预期的治疗效应。这种治疗效应源于一次新的抽样 δ_0，以及约为 τ 的新偏移。

式（1）中的分解在结构上与随机效应荟萃分析中研究的参数相似；在随机效应荟萃分析中，假定研究特异性治疗效应是从某个分布（通常为正态分布）中抽取的独立同分布样本。在本研究中，我们定义并估计一组参数，它们在保持式（1）这一熟悉形式的同时，还具有精确的因果解释。为此，我们在潜在结果框架中纳入了 δ_s 所体现的研究特异性异质性。具体而言，我们将 δ_s 所代表的变异作为额外参数纳入标准潜在结果符号中。

2.1 数据

在上述示例的基础上补充符号，假设我们拥有一组 m 试验的个体参与者数据，这些试验均考察了同一组治疗。再次使用集合 $\mathcal{S} = \{1, ..., m\}$ 为这些试验编索引，并用集合 $\mathcal{A}$表示治疗。对于给定试验 $s \in \mathcal{S}$ 中的每位参与者 i，若该试验共有 n_s 名参与者，我们有结局数据 Y_i、基线协变量信息 X_i 和治疗分配 A_i，其中 $A_i \in \mathcal{A}$。我们还假设，拥有来自目标人群中个体随机样本的基线协变量信息；该样本不需要治疗或结局信息。与 Dahabreh 等人（2022）一样，我们令 $S = 0$ 表示目标人群中的个体，并引入另一个变量 R：对于试验集合中的参与者，该变量取值为 1；对于目标人群中的成员，则取值为 0。因此，完整数据集中的每位个体的观测数据——也就是试验集合和目标人群两部分数据的合并——形式如下:

$$(R_i Y_i, R_i A_i, X_i, S_i).$$

$R_i Y_i$ 和 $R_i A_i$ 在目标人群中均为零，表明这类数据不可用。目标人群样本与所有试验样本的总样本量为 $n = n_0 + n_1 + \cdots + n_m$。

2.2 提出新因果量的理由

在介绍因果可解释荟萃分析时，Dahabreh 等人（2022）定义了取决于分配的治疗 $a \in \mathcal{A}$ 的潜在结局 $Y(a)$。不同的因果估计目标描述了这些潜在结局在不同目标人群中的分布。例如， $E[Y(a)|R=0]$ 表示目标人群中接受治疗 a 时潜在结局的期望值。为了识别诸如 $E[Y(a)|R=0]$这样的量，Dahabreh 等人假设在给定基线协变量的条件下，不同 S取值对应的潜在结局具有可交换性。他们还表明，在较弱的均值可交换性假设下， $E[Y(a)|R=0]$ 仍可识别。均值可交换性大致定义为以下假设：

$$E[Y(a)|X=x, S=0] = E[Y(a)|X=x, S=s] \tag{2}$$

对于 $s=1$，……， m，并附加条件，以确保该假设仅适用于S 取值中协变量模式 $X=x$ 出现概率大于零的情况。

直观而言，在假设（2）实际上并不成立时仍作此假设，主要风险在于各项试验特有的细节可能会干扰基于合并数据估计的迁移治疗效应。如果荟萃分析集合中的试验在样本量上差异很大，这个问题尤其突出。在这种情况下，违反（2）可能导致因果估计值过度偏重样本量最大的试验所处的特定条件。这种加权方式没有临床意义，只是均值可交换性假设造成的产物。

之所以要参照此类研究间异质性，直接源于标准随机效应荟萃分析背后的逻辑（Higgins、Thompson 和 Spiegelhalter，2009）。在这一范式下，对一组试验进行荟萃分析时，假定关注的底层参数（通常为平均治疗效应）在不同试验间存在差异。更确切地说，在传统随机效应模型中，假定每项研究的潜在真实治疗效应均从某个效应分布中随机抽取。该底层分布的方差往往值得关注，并用于量化研究间异质性的程度。

我们所提方法将从研究 $s \in \mathcal{S}$ 迁移至目标人群的平均效应建模为围绕总体均值变化，其变化由潜在随机变量的抽样所决定。该潜在随机变量的抽样将这些潜在的平均效应固定在不同数值上，而这些数值之间的差异反映了超出协变量差异所导致的研究间异质性。这些方法与传统随机效应模型之间的关键区别在于：潜在参数值的异质性（例如，研究间治疗效应的异质性）源于研究参与者潜在结局的变异，因此保留了标准荟萃分析所缺乏的因果解释。

请回忆一下，促成本研究的一个关键因果问题是：如果某个个体在研究 $a \in \mathcal{A}$ 中接受分配的治疗 $s \in \mathcal{S}$，我们预期会出现什么结局？这个问题表明，有必要区分源自一个试验情境与源自另一个试验情境的潜在结局。采用 VanderWeele 和 Hernán（2013）的符号，我们将一名受试者在研究$s \in \mathcal{S}$ 所处条件下接受治疗 $a \in \mathcal{A}$ 时的潜在结果写成具有两个自变量：一个用于指定治疗分配，另一个用于固定特定环境下的现有条件，例如研究 s中的治疗组 a：

$$Y(a, k_s^a).$$

这里， k_s^a是随机变量 $K_s^a(a)$的一次实现。由于我们会为未在研究 s中接受治疗 a 的个体设想 $K_s^a(a)$的取值，因此其记号仿照标准潜在结局，因为个体在某项具体研究中的经历可能取决于其接受的治疗。我们假设，每一种研究与治疗组的组合都对应一个这样的随机变量；对于给定治疗 a，这些随机变量在各项研究中构成来自某个共同分布的独立同分布样本：

$$K_1^a(a), ..., K_m^a(a) \overset{i.i.d.}{\sim} F_{K^a(a)}.$$

我们不对 $K_s^a(a)$作特定解释，只要求它反映个体若处于研究 s的治疗组 a 中本会经历的条件。也就是说，具有相同该随机变量取值的两个个体，可以理解为在等同条件下被分配到治疗 a。同样，我们将 $K_s^a(a)$视为一种可能的反事实变量，即使对于未被分配到治疗 a 或未参与研究 s的个体，它也能确定环境/治疗组组合中的异质性。

我们假设这些随机变量是来自某个共同分布的独立同分布样本，并不是因为我们认为这能最准确地近似诱发研究间异质性的数据生成过程。相反，在缺乏额外信息的情况下，这一假设是我们对这种异质性本质所能作出的最佳判断。对于只能获取结局、基线协变量和治疗组信息的（荟萃分析）分析者而言，残余的研究间异质性就如同独立同分布的噪声，其作用类似于简单线性回归模型中的残差。如果我们掌握了更多关于试验间异质性本质的信息，就可以放宽或改进这一假设。

这种异质性的影响与试验参与本身的影响有关，但并不相同；Dahabreh 等人（2019）研究了后者。在此，我们构建了一个概念框架，用于理解参与某项试验而非另一项试验的影响，而不是参与试验本身的影响。这项工作在一定程度上可以

被视为对早期因果可解释荟萃分析相关研究中所定义的因果估计目标的阐释作进一步澄清。鉴于试验间异质性的影响，我们研究参与正在研究的某项试验，或参与与正在研究的试验相似的试验所产生的影响，而非目标人群中治疗分配所产生的更一般性影响。

2.3 定义因果估计目标

与 Dahabreh 等人（2022）一样，我们关注的是如何将来自一系列试验的推断迁移到目标人群。由于 $K_s^a(a) = k_s^a$ 的实际取值刻画了由试验 s 引起的潜在结局异质性，因此，若要评估目标人群成员参与试验 s 时可能观察到的结局类型，与之相关的因果量是

$$\mu_{a,0}(k_s^a) = E[Y(a, k_s^a)|R = 0]. \tag{3}$$

这是一个固定的因果量，可解释为目标人群中的平均潜在结局，其中治疗 a 应用方式的异质性固定为与试验 s 相关联的取值。

我们认识到，政策制定者和临床医生通常最关注其目标人群中的治疗效应，而不只是平均潜在结局。为使本文的阐述更为简明，我们将重点放在平均潜在结局上。不过，也可以定义因果量（如（3））的对比，例如

$$\tau_{a,a'}^s = \mu_{a,0}(k_s^a) - \mu_{a',0}(k_s^{a'}) = E[Y(a, k_s^a)|R = 0] - E[Y(a', k_s^{a'})|R = 0]$$

可以定义与试验 s 条件下目标人群中的平均治疗效应相对应的对比。关注此类效应也使我们能够放宽对平均潜在结局的均值可交换性假设，并以平均效应可交换性取而代之。

回到式 (3)，请注意，对于每项研究，其取值实现后构成这些迁移的潜在结局的随机变量，是潜在随机样本 $K_1^a(a), ..., K_m^a(a)$ 的函数：

$$\mu_{a,0}(a, K_1^a(a)), ..., \mu_{a,0}(a, K_m^a(a)) \tag{4}$$

因此，它们也构成来自某个分布的简单随机样本。

我们假设这个随机样本 $\mu_{a,0}(K_1^a(a)), ..., \mu_{a,0}(K_m^a(a))$ 来自一个均值为 $\mu_{a,0}$且具有有限方差的分布。在不作任何其他建模假设的情况下，我们可以将样本中的每个此类随机变量分解为

如下形式：

$$\mu_{a,0}\left(K_s^a(a)\right) = \mu_{a,0} + \Delta_{0,s}^a, \tag{5}$$

其中$\Delta_{0,\,s}^a$ 是一个均值为零且具有有限方差的随机变量。请注意，(5) 并未对$\mu_{a,\,0}\left(K_s^a(a)\right)$施加任何特定的建模假设；它只是将$K_s^a(a)$的期望值周围的随机变异标记为$\Delta_{0,\,s}^a$。重新排列 (5)，可得

$$\begin{aligned}\Delta_{0,s}^a &= \mu_{a,0}\left(K_s^a(a)\right) - \mu_{a,0} \\ &= \mu_{a,0}\left(K_s^a(a)\right) - E_{K^a(a)}\left[\mu_{a,0}\left(K_s^a(a)\right)\right].\end{aligned}$$

(5) 中的分解有助于我们更好地理解潜在随机变量 $K_s^a(a)$在造成不同研究之间因果量的系统性差异时所发挥的作用。具体而言，抽取一个 $K_s^a(a) = k_s^a$，就对应于$\Delta_{0,\,s}^a = \delta_{0,\,s}^a$ 的一个实现值；这又会使从研究 s 迁移至目标人群的 a下的平均潜在结果发生偏移。

回想一下我们在方程 (1) 中对方法所作的简化表述，因此，如果目标人群中的一名成员在研究 s 中被分配接受治疗 a 而非 a'，我们会观察到的治疗效应模型为

$$\begin{aligned}\mu_{a,0}(k_s^a) - \mu_{a',0}(k_s^{a'}) &= \mu_{a,0} + \delta_{0,s}^a - \left(\mu_{a',0} - \delta_{0,s}^{a'}\right) \\ &= \left(\mu_{a,0} - \mu_{a',0}\right) + \left(\delta_{0,s}^a - \delta_{0,s}^{a'}\right).\end{aligned} \tag{6}$$

我们可以将$\tau_{a,\,a',\,0} = \mu_{a,\,0} - \mu_{a',\,0}$ 视为目标人群中治疗 a 相对于 a' 的“总体平均效应”，而将$\delta_{0,\,s}^a - \delta_0^{a'}$，$_s$ 视为从研究 s所在环境进行迁移时，该总体效应的异质性。在方程 (1) 中，这两个量分别对应于 τ 和 δ_s。

除了考虑目标人群中的个体参与我们所收集的某项试验时的预期潜在结果，我们还可以问：如果开展一项新试验，并从目标人群中招募一个简单随机样本，我们会预期出现怎样的潜在结果？这里我们作出一项重要假设：在所收集的试验中导致预期潜在结果存在异质性的同一随机过程，也适用于一项采用相同治疗的新试验。

我们可以在原来的随机样本中添加 $K_0^a(a)$ 来表示这一假设：

$$K_1^a(a), ..., K_m^a(a), K_0^a(a) \overset{i.i.d.}{\sim} F_{K^a(a)}$$

并像之前一样定义 $\mu_{a,\,0}(k_0^a)$ 和 $\mu_{a,\,0}\left(K_0^a(a)\right)$。

这一框架类似于随机效应荟萃分析中的一种观点：我们可以对未纳入荟萃分析的研究的治疗效应作出推断（Higgins、Thompson 和 Spiegelhalter，2009）。同样，我们关于 $K_0^a(a)$ 的独立同分布假设反映了我们对支配研究间异质性的随机过程所作的最佳判断。在缺乏额外信息的情况下，我们假定，试验样本中随机对照试验之间产生系统性差异的异质性来源，同样适用于目标人群中的假设性随机对照试验；并且，这种额外变异源自共同分布中的独立同分布抽样。如下文所述，我们在构建预测区间时采用这一假设，使区间以某个指定概率包含 $\mu_{a,\ 0}(k_0^a)$。

估计未观测试验中的结局有几种可能的动机。其一是将估计结果用于规划和筹备新的临床研究。临床医生可能希望估计，在一项计划开展的研究中，若参与者与目标人群成员相似，他们可能会预期观察到怎样的治疗效应。或者，临床试验中的迁移估计值可以作为一种界限，帮助我们判断在实际中可能观察到的效应。估计未观测试验中的结局也能得出这样的界限，不过它不会过度受试验集合中任何特定研究所具有的个别特性的影响。

3 因果量的识别与估计

3.1 识别

我们首先考虑 $\mu_{a,\ 0}(k_s^a)$ 的识别，它表示如果目标人群中的一名成员在研究 s 中被分配接受治疗 a，本可观察到的预期结局。为了将其表示为观测数据的泛函，我们作出以下识别假设，其中许多与 Dahabreh 等人（2022）以及 VanderWeele 和 Hernán（2013）提出的假设相似：

1. **试验间均值可交换性：**

$$E[Y(a,k_s^a)|X=x,S=s_1]=E[Y(a,k_s^a)|X=x,S=s_2]$$

对于所有 s_1、$s_2\in\{0、1、.....、m\}$ 以及满足 $f(x、S=s_1)\neq 0$ 和 $f(x、S=s_2)\neq 0.$ 的 $x\in\mathcal{X}$，注意，此处试验 s_1 和 s_2 中的异质性在两个环境中均固定为 k_s^a。

2. **试验内治疗组间可交换性：**

$$Y(a,k_s^a)\perp\!\!\!\perp A|(X,S=s)$$

对于所有 $a \in \mathcal{A}$、 $k_s^a \in \mathcal{K}^a$ 和 $s \in \mathcal{S}$。

3. **一致性：** 如果对于个体 i， $A_i = a$ 和 $K_i^a(a) = k^a$，那么 $Y_i(a, k^a) = Y_i$。

4. **正值性：** 对于所有 $s \in \mathcal{S}$，如果 $f(x, R = 0) \neq 0$，那么 $P(S = s|X = x) > 0$。

5. **()的分布：** 对于 $s = 1, ..., m$ 和 $a \in \mathcal{A}$，每个随机变量 $K_s^a(a)$ 均构成从分布 $F_{K^a(a)}$中进行的一次独立同分布抽样。

6. **()的恒定性（按治疗组/研究）：** 在每个治疗组 a 和研究 s内，每位参与者的 $K^a(a)$ 值固定为 $K_s^a(a) = k_s^a$的实现值。也就是说，如果 $A = a$ 和$S = s$，那么

$$K_i^a(a) = K_s^a(a) = k_s^a$$

对于 $i \in 1, ..., n_s$。

假设 3 和 5 涉及对符号的轻微滥用，即我们定义随机变量 $K_i^a(a)$，表示个体 i所接受的治疗版本。这里的关键在于，对于试验 s中的所有参与者，该随机变量均取 $K_s^a(a)$ 的同一个实现值。另一种做法是，假设存在一个层级模型，其中治疗组 a 和研究 s 中的每个个体都对应一个随机变量 $(K_s^a(a))_i, i = 1, ..., n_s$，其中心位于 $K_s^a(a) = k_s^a$的实现值处。本文着重讨论试验/治疗组层面的异质性，以大大简化数学表述。然而，在某些假设下，上述类型的层级模型可以得出本文所呈现的相同识别结果。

在这些假设下，我们识别出 $\mu_{a,\ 0}(k_s^a)$，其形式为

$$\psi_{s,0}(a) = E[E[Y|X, S = s, A = a]|R = 0]. \tag{7}$$

也就是说，我们对研究 s 中治疗 a 条件下结局与协变量 X 之间关系的回归函数进行平均，所依据的是观测到的目标人群中此类协变量的分布。（该结果的完整证明见附录。）这一观测数据泛函类似于 Dahabreh 等人（2022）中定理 1 的公式（6）所识别的泛函，只是他们使用的 $R = 1$ 在我们的情形中被 $S = s$ 取代。

3.2 估计

令 $g_a^s(X) = E[Y|X, S = s, A = a]$，我们可以采用结局模型/标准化方法，对 $E[Y|X, S = s, A = a]$ 的估计值 $\hat{}g_a^s(X)$ 按目标

人群:

$$\hat{\psi}_{s,0}(a) = \left\{\sum_{i=1}^{n} I(S_i = 0)\right\}^{-1} \sum_{i=1}^{n} I(S_i = 0)\left\{\hat{g}_a^s(X_i)\right\}$$

另一种方法是逆概率加权，即根据研究 s 中的参与者与目标人群成员的相似程度，对观察到的结局进行加权。采用此方法的估计量为

$$\hat{\psi}_{s,0}^{ipw}(a) = \left\{\sum_{i=1}^{n} I(S_i = 0)\right\}^{-1} \sum_{i=1}^{n} \left(\frac{I(A_i = a)}{\hat{e}_a^s(X_i)}\right) I(S_i = s)\frac{\hat{p}_0(X_i)}{\hat{p}_s(X_i)} Y_i$$

其中，$\hat{p}_0(X_i)$ 估计 $P(R = 0|X_i)$，$\hat{e}_a(X_i)$ 估计 $P(A_i = a|S = s,\ X_i)$，而 $\hat{p}_s(X_i)$ 估计 $P(S = s|X_i)$。从 (7) 推导这一结果的识别过程见附录。

最后一种方法是所谓的增广逆概率加权（AIPW）估计量，它结合了上文介绍的两种方法:

$$\hat{\psi}_{s,0}^{aipw}(a) = \left\{\sum_{i=1}^{n} I(S_i = 0)\right\}^{-1} \sum_{i=1}^{n} \left[I(S_i = 0)\left\{\hat{g}_a^s(X_i)\right\} + \left(\frac{I(A_i = a)}{\hat{e}_a^s(X_i)}\right) I(S_i = s)\frac{\hat{p}(X_i)}{1 - \hat{p}(X_i)}(\hat{g}_a^s(X_i) - Y_i)\right]$$

这种方法通过对研究 s 中组别 a 的结局模型残差进行逆概率加权，为结局模型方法增加了一项校正。

我们感兴趣的第二个因果量是 $E[Y(a,\ k_0^a)|R = 0]$，其中 k_0^a 是前一节所述随机变量 $K_0^a(a)$ 的实现值。在缺乏关于分布 $F_{K^a(a)}$ 的额外信息时，我们对 $E[Y(a,\ k_0^a)|R = 0]$ 的估计策略依赖于以下近似，其完整细节见附录:

$$\begin{aligned}
E[Y(a, K_0^a(a))|R = 0] &= \sum_{k^a \in \mathcal{K}^a} E\left\{E[Y(a, K_0^a(a))|R = 0, K_0^a(a) = k^a]|R = 0\right\} P(K_0^a(a) = k^a) \\
&\approx \sum_{k^a \in \mathcal{K}^a} E\left\{E[Y(a, K_0^a(a))|R = 0, K_0^a(a) = k^a]|R = 0\right\} \left\{\frac{1}{m}\sum_{s=1}^{m} \mathbb{1}(k^a = k_s^a)\right\} \\
&= \frac{1}{m}\sum_{s=1}^{m} E[Y(a, k_s^a)|R = 0] \\
&= \frac{1}{m}\sum_{s=1}^{m} E[E[Y|X, S = s, A = a]|R = 0] \qquad \text{by (7)} \\
&= \frac{1}{m}\sum_{s=1}^{m} \psi_{s,0}(a).
\end{aligned}$$

再次，对上述量在不同治疗下取对比值，即可识别

我们感兴趣的治疗效应，例如式（6）所示形式的因果量。随后，我们可以使用上述任意一种方法，分别估计各个 $\psi_{s,0}(a)$，从而估计 $\frac{1}{m}\sum_{s=1}^{m}\psi_{s,\ 0}(a)$。例如，我们可以使用以下方法估计该量

$$\frac{1}{m}\sum_{s=1}^{m}\hat{\psi}_{s,0}(a)=\frac{1}{m}\sum_{s=1}^{m}\left(\left\{\sum_{i=1}^{n}I(S_i=0)\right\}^{-1}\sum_{i=1}^{n}\{\hat{g}_a^s(X_i)\}\right).$$

3.3 研究间变异性的估计

传统随机效应荟萃分析中的一个核心关注参数是研究间方差，它描述了各项研究背后效应的变异性。回顾式（5）中的分解，我们研究中的对应参数是 $\Delta_{0,s}^{a}$ 的方差，它刻画了迁移后的平均潜在结局在研究间的变异性。我们提出一种方差估计方法，其灵感来自 Rao 等人（1981）推导的矩估计法，随后 DerSimonian 和 Laird（1986）在其关于随机效应荟萃分析的开创性研究中应用了该方法。我们的估计值与传统随机效应估计值的关键区别在于，实际中，$\mu_{a,0}(k_1^a)...,\mu_{a,0}(k_m^a)$ 的估计值彼此相关，因为它们依赖于目标人群的同一个样本。这种相关性使所得估计量的推导及其形式更加复杂。

在下面的推导中，我们在一个高度一般化的环境中考虑研究间变异性的估计；与因果框架唯一相关的是研究特异性估计值之间的相关性。在一个简化的环境中，假设我们有 m 个研究特异性均值 $\mu_1,...,\mu_m$，它们来自一个均值为 μ 、方差为 γ^2 的分布 F_μ。通过令 $\mu_s=\mu_{a,0}(k_s^a)$，即可得到与我们环境的类比。我们用 $\hat{}\mu_1,...,\hat{}\mu_m$ 估计这些均值，其中每个估计值对于相应的研究特异性均值而言都是无偏的。与我们对应的量为 $\hat{}\mu_s=\hat{}\psi_{s,0}(a)$。我们使用研究特异性估计值 $\hat{}\mu=\sum_{s=1}^{m}w_s\hat{}\mu_s$ 的加权和来估计总体均值 μ，其中 $\sum_{s=1}^{m}w_s=1$。（例如，简单取平均时，我们可能有 $w_s=\frac{1}{m}$。）此外，令 $\mathrm{Var}(\hat{}\mu_s|\,S=s)=\sigma_s^2$ 表示每个研究特异性估计量的抽样方差，并令 $\mathrm{Var}(\hat{}\mu_s)=\sigma_s^2+\gamma^2$，它同时反映研究内方差和研究间方差。令 $Q=\sum_{s=1}^{m}(\hat{}\mu_s-\hat{}\mu)^2$ 和 $C_s=-2\sum_{i\neq s}w_i(1-w_s)\sigma_{is}+\sum_{i\neq s}\left[\sum_{j\neq i,j\neq s}w_iw_j\sigma_{ij}\right]$，我们可以证明

$$E[Q]=m\sum_{s=1}^{m}w_s^2(\sigma_s^2+\gamma^2)+\sum_{s=1}^{m}\left[(1-2w_s)(\sigma_s^2+\gamma^2)+C_s\right].$$

γ^2 的基于矩的估计量随后由满足以下条件的值 $\tilde{\gamma}^2$ 给出

$$\sum_{s=1}^{m}(\hat{\mu}_s-\hat{\mu})^2=m\sum_{s=1}^{m}w_s^2(\sigma_s^2+\tilde{\gamma}^2)+\sum_{s=1}^{m}\left[(1-2w_s)(\sigma_s^2+\tilde{\gamma}^2)+C_s\right].$$

求解这样的 $\tilde{\gamma}^2$，可得

$$\tilde{\gamma}^2 = \frac{\sum_{s=1}^{m} (\hat{\mu}_s - \hat{\mu})^2 - \sum_{s=1}^{m} \sigma_s^2 \left\{ m w_s^2 + (1 - 2w_s) \right\} - \sum_{s=1}^{m} C_s}{\sum_{s=1}^{m} \left\{ m w_s^2 + (1 - 2w_s) \right\}}.$$

因此，我们对研究间异质性的最终估计值为

$$\hat{\gamma}^2 = \max(0, \tilde{\gamma}^2). \tag{8}$$

上述估计量的完整推导见附录。我们的估计量与 Rao 等人（1981年）提出的传统随机效应估计量之间的关键区别，在于 $\hat{\gamma}^2$ 的分子中包含了 $-\sum_{s=1}^{m} C_s$。这样做可以“校正”传统估计量，以考虑研究特定估计值之间的非零相关性。除了我们所处的环境外，该估计量通常也适用于任何通过对相关量取加权平均来估计某个潜在总体均值的分析；不过请注意，当个体估计值本身存在偏倚时，该估计量也会产生偏倚。

3.4 对新试验结局的推断

对于试验人群成员参加一项已观察到的试验时我们预期会出现的结局，例如 $\mu_{a,0}(k_s^a)$，$s \in \mathcal{S}$，可以通过多种方式进行推断，包括对目标人群和研究 s 的数据进行简单自助重抽样，或基于 $\hat{\psi}_{s,0}(a)$ 的抽样分布进行渐近近似。此处，我们重点讨论对 $\mu_{a,0}(K_0^a(a))$ 的推断。在这一估计目标中，我们将 $K_0^a(a)$ 视为随机变量，因为它指的是一项我们尚未观察到的试验中的结局。因此，对所有已观察到的试验和目标人群应用简单自助法，无法捕捉开展新试验所带来的额外变异。也就是说，我们的预测区间旨在同时涵盖估计量 $\frac{1}{m}\sum_{s=1}^{m} \hat{\psi}_{s,0}(a)$ 的抽样变异（如观测数据所反映的那样），以及关于未观测试验的额外不确定性。假设未观测试验面临的异质性与已观察到的试验之间的异质性相似，我们便可以进一步开展分析。此处，我们提出三种构建这些预测区间的方法，并将在后文通过模拟实验对其进行评估。

—

3.4.1 基于$\hat{\gamma}^2$的推断

上文中，我们推导出了估计量 $\hat{\gamma}^2$，通常将其解释为测量协变量所引起的部分之外的研究间异质性的估计值。当我们将 $\hat{\mu}_s$ 替换为 $\hat{\psi}_{s,0}(a)$，并将 $\hat{\mu}$ 替换为 $\frac{1}{m}\sum_{s=1}^{m} \hat{\psi}_{s,0}(a)$ 时，$\hat{\gamma}^2$ 便估计 $\text{Var}\left(\mu_{a,0}(\underline{K}_s^a(a))\right)$。也就是说，在我们的框架中，$\hat{\gamma}^2$ 量化了不同试验间迁移的潜在结局期望值的变异性，而这种变异性由环境特定

异质性驱动。为了构建一个以 $\frac{1}{m}\sum_{s=1}^{m}\hat{\psi}_{s,0}(a)$ 为中心的预测区间，我们采用 Higgins等人（2009）在推断“未指定研究中的效应”时所考虑的同一形式的区间：

$$\frac{1}{m}\sum_{s=1}^{m}\hat{\psi}_{s,0}(a) \pm t_{m-2}^{\alpha}\sqrt{\hat{\gamma}^2 + \widehat{\mathrm{Var}}\left(\frac{1}{m}\sum_{s=1}^{m}\hat{\psi}_{s,0}(a)\right)}, \tag{9}$$

其中，$\widehat{\mathrm{Var}}\left(\frac{1}{m}\sum_{s=1}^{m}\hat{\psi}_{s,0}(a)\right)$ 是通过对所有已观测研究和目标人群应用简单自助法获得的，而 t_{m-2}^{α} 是具有 $m-2$ 自由度的 t 分布的第（$1-\alpha$）分位数。

3.4.2 基于自助法的推断

近似我们的观测估计值与未观测试验估计值之间差异的一种方法，是使用 $m-1$ 个试验子集的数据计算 $\frac{1}{m-1}\sum_{s=1}^{m-1}\hat{\psi}_{s,0}(a)$，并将该结果与第 m个被留出的试验的迁移的潜在结局进行比较（对所有试验重复此过程）。该程序旨在捕捉与试验间异质性相关的变异。为进一步捕捉观测数据中的抽样变异，我们将留一法与简单自助法或野生自助法结合使用。下面将介绍这两种自助法。两种程序的结果都是构建$\mu_{a,0}(K_s^a(a))$分布的估计值。为简化下文中的记号，令 $\hat{\mu}_0 = \frac{1}{m}\sum_{s=1}^{m}\hat{\psi}_{s,0}(a)$和 $\hat{\mu}_s = \hat{\psi}_{s,0}(a)$。

令 D_s表示我们从试验 $s=1,...,m$的原始样本中获得的所有数据。令 X 表示从目标人群中获得的协变量数据。对于 $b=1,...,B$,

1. 从 m 项研究中随机选择一项，将其视为我们要对其进行预测的“未观测”试验。将这一选择记为 $s_b \in \{1,...,m\}$。2. 从 D_{s_b} 中抽取一个简单自助样本 $D_{s_b}^*$，并从 X 中抽取一个简单自助样本 X_1^*，然后利用这两个数据集估计 $\hat{\mu}_{s_b}^*$。3. 从其他试验数据中抽取简单自助样本 $\{D_s^*: s \neq s_b\}$。再从 X中抽取另一个自助样本X_2^*。利用这些样本计算 $\mu^* = \frac{1}{m-1}\sum_{s\neq s_b}\hat{\mu}_{s\circ}^*$ 。4. 计算估计残差 $\delta_b^* = \mu^* - \hat{\mu}_{s_b}^*$，以量化我们的预测误差。5. 最后，利用一组新的自助样本，对所有原始数据（包括在计算预测误差时留出的试验）计算另一个估计值 $\hat{\mu}^*$。令 $\hat{\mu}_b^{pred} = \hat{\mu}^* - \hat{\delta}_b^*$。

我们还考虑使用基于估计量 $\hat{\mu}_s$ 的影响函数的野生自助法来构造预测区间。该过程的步骤与上述简单自助法类似，不同之处在于，Matsouaka 等人（2022）的方法不再对数据进行有放回抽样。请注意，影响函数的具体形式

取决于估计量。由于我们在下文所述的模拟研究中使用结局模型，因此在实施野生自助法时使用该模型的影响函数。有关影响函数定义和推导的更多细节，可参见 Tsiatis（2006）等文献。

4 模拟研究

为了评估上述三种方法的性能，我们按如下方式开展模拟研究 s.

4.1 数据生成过程

协变量：假设每个 m 试验和目标人群样本中都有三个协变量 X_1、X_2、X_3。每项试验共有 100 名参与者，分为两个治疗组；目标人群样本量为 1000。每种环境下的协变量 $s = 0, 1 \ldots m$ 按如下方式生成：

1. 通过在区间 $[0, 1.5]$ 内生成 m 个等间距值，并将 $\boldsymbol{\mu}_s$ 设为其中第 s 个值，确定研究 s 中的平均协变量向量 $\boldsymbol{\mu}_s \in \mathbb{R}^3$（对 $\boldsymbol{\mu}_s$ 中的每个条目重复该值三次）。因此，若 $m = 3$，则 $\boldsymbol{\mu}_1 = (0, 0, 0)^T$、$\boldsymbol{\mu}_2 = (0.75, 0.75, 0.75)^T$，以及 $\boldsymbol{\mu}_3 = (1.5, 1.5, 1.5)^T$。在每种情景中，目标人群的平均协变量向量为 $\boldsymbol{\mu}_0 = (1, 1, 1)^T$。

2. 对研究 s 中的每位 n_s 参与者，抽取一个向量 $(\mathbf{X}s)_i \sim \mathcal{N}(\boldsymbol{\mu}_s, \Sigma)$，其中

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}.$$

请注意，这个矩阵在各项研究和目标人群中均相同。

潜在结局：有了这些协变量值后，我们将研究 s 中每位参与者在接受治疗 a 时的潜在结局生成为 $(X_{s,1}, X_{s,2}, X_{s,3})_i^T$ 的线性组合。也就是说，

$$(Y(a, k_s^a) | S = s, \mathbf{X} = \mathbf{x})_i = \beta_0 + \beta_1 x_{s,1} + \beta_2 x_{s,2} + \beta_3 x_{s,3} + \delta_s^a + \epsilon, \tag{10}$$

其中 $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0.5$ 和 $\epsilon \sim N(0, 1)$。值 δ_s^a 是研究 s 中实施治疗 a 所导致的潜在结果中随机变异的实现值；它在研究 s 的所有参与者中均保持不变。这种变异的分布在不同模拟情景中有所不同，详情如下。

根据（10）中的设定，感兴趣的估计目标的真实值为

$$\begin{aligned} E[Y(a,k_0^a)|R=0] = E[E[Y(a,k_0^a)|S=0,\mathbf{X}]] &= 0.5 + 0.5E[X_{0,1}] + 0.5E[X_{0,2}] + 0.5E[X_{0,3}] + \delta_0^a + E[\epsilon] \\ &= 0.5 + 0.5(1) + 0.5(1) + 0.5(1) + \delta_0^a + 0 \\ &= 2 + \delta_0^a. \end{aligned} \tag{11}$$

这是目标人群成员若参与一项“未观察到的”试验时，在治疗 a 下的真实期望值。每项已观察到的试验中，受试者以 1:1 随机化的方式分配接受治疗 a 或安慰剂。

估计：使用前文定义的符号，我们利用正确设定的结局模型估计研究 s 中治疗 a下的平均观测结局（记为 $g_a^s(x) = E[Y|X=x,S=s,A=a]$）

$$\hat{g}_a^s(x) = \mathbf{x}^T\hat{\boldsymbol{\beta}}_s,$$

其中，$\mathbf{x}^T = (1, x_1, x_2, x_3)^T$，而 $\boldsymbol{\beta}_s = (\beta_{s,0}, \beta_{s,1}, \beta_{s,2}, \beta_{s,3})^T$。将该结局在目标人群协变量分布上取平均后的期望值（记为 $\psi_{s,0}(a) = E[E[Y|X,S=s,A=a]|R=0]$）估计为

$$\hat{\psi}_{s,0}(a) = \frac{1}{n_0}\sum_{i:S_i=0}\hat{g}_a^s(X_i).$$

如上所述，各个研究中 $\psi_{s,0}(a)$ 的平均值，是我们对目标估计目标$\mu_{a,0}(k_0^a)$ 的最佳估计：

$$\hat{\mu}_{a,0}(k_0^a) = \frac{1}{m}\sum_{s=1}^{m}\hat{\psi}_{s,0}(a).$$

随后，我们使用上述三种方法，构造一个以 $\hat{\mu}_{a,0}$ 为中心的预测区间，旨在以某个预先指定的概率包含 $E[Y(a,k_0^a)|R=0] = 2 + \delta_0^a$。

4.2 模拟情景

上述数据生成过程经过了高度简化。它假设所有情景中的真实结局模型完全相同，并且我们正确设定了该模型。未来的工作将使这一设置更加复杂，以探究其他问题，例如模型设定错误。目前，这个简单情景让我们聚焦于两个主要参数，并在模拟中改变它们：

1. 荟萃分析研究数据集中试验的数量。该数量取值范围包括 m =5、10、15、30 和 50 项研究。虽然我们很少会预期荟萃分析中有 50 项研究，但该情景的目的是

探究随着我们从越来越多的研究中收集数据，某种给定方法是否会出现覆盖不足或覆盖过度的情况。
2. 特定设置的变异分布。设置包括：、均匀分布$[-2, 2]$ $N(0,\ 1)$、指数分布(1)−1，以及帕累托分布(1，3)。

上述环境定义了 $5 \times 4 = 20$ 种不同的模拟情景。我们使用根据上述数据生成过程模拟的 1000 个人工数据集评估每种情景。在适用情况下，我们的估计程序中任何自助法的应用均包含 1000 次重复。对于所关注的某种预测区间构造方法，我们将覆盖概率估计为这 1000 次迭代中包含真实参数值的预测区间所占的比例。

4.3 结果

图1展示了根据自助法估计值的分位数构造的预测区间的覆盖率估计值。第二组图则使用正态分布近似，即自助法估计值的均值 ± 1.96 乘以自助法估计值的标准差。在图中，矩估计法表示使用我们对 $\hat{}\gamma^2$ 的估计值构造的区间，如式（9）所示。（请注意，两幅图中的矩估计法结果相同，因为它们不受此类选择的影响。）

总体而言，在不同残余异质性分布下，基于研究间方差校正估计量的矩估计法区间表现优于基于自助法的替代方法。自助法的正态近似也优于基于经验分位数的预测区间，不过其表现仍不及矩估计法估计量。尽管假设荟萃分析研究数量极大的模拟有助于理解我们的方法在渐近情况下的某些特性，但我们认识到，15 项或更多研究的数量远超绝大多数荟萃分析数据集中的研究数量，尤其是考虑到还需要个体参与者数据。因此，这些模拟结果表明，在研究人员希望估计从目标人群中招募的某项新的、未观测试验中的效应时，大多数情况下应采用基于研究间方差校正估计量的矩估计法区间。

尽管两种自助法替代方法的表现相近，但值得注意的是，野生自助法在计算方面具有不可忽视的优势。在配备英特尔酷睿 i7-9700 处理器的标准 Windows 计算机上，在研究数量为 5、样本量如上所述的情形下，野生自助法生成预测区间的速度是简单自助法的两倍多。在这类情形下，绝对耗时都微不足道（16 秒对 6 秒）；然而，随着研究数量增加，且目标人群样本量大幅增大，野生自助法在计算方面会变得更具吸引力。

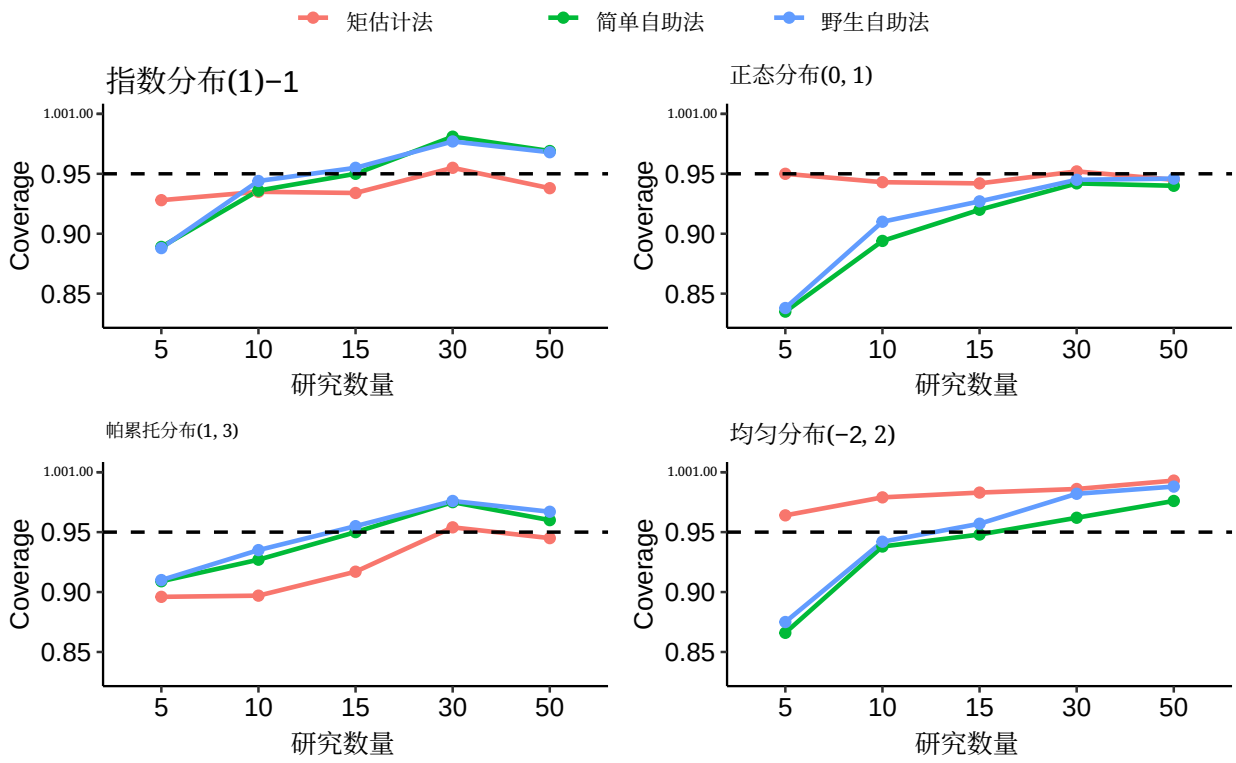


图1：根据自助法估计值的分位数构建的预测区间的覆盖率。每个图对应特定设置的变异的一种单独分布。

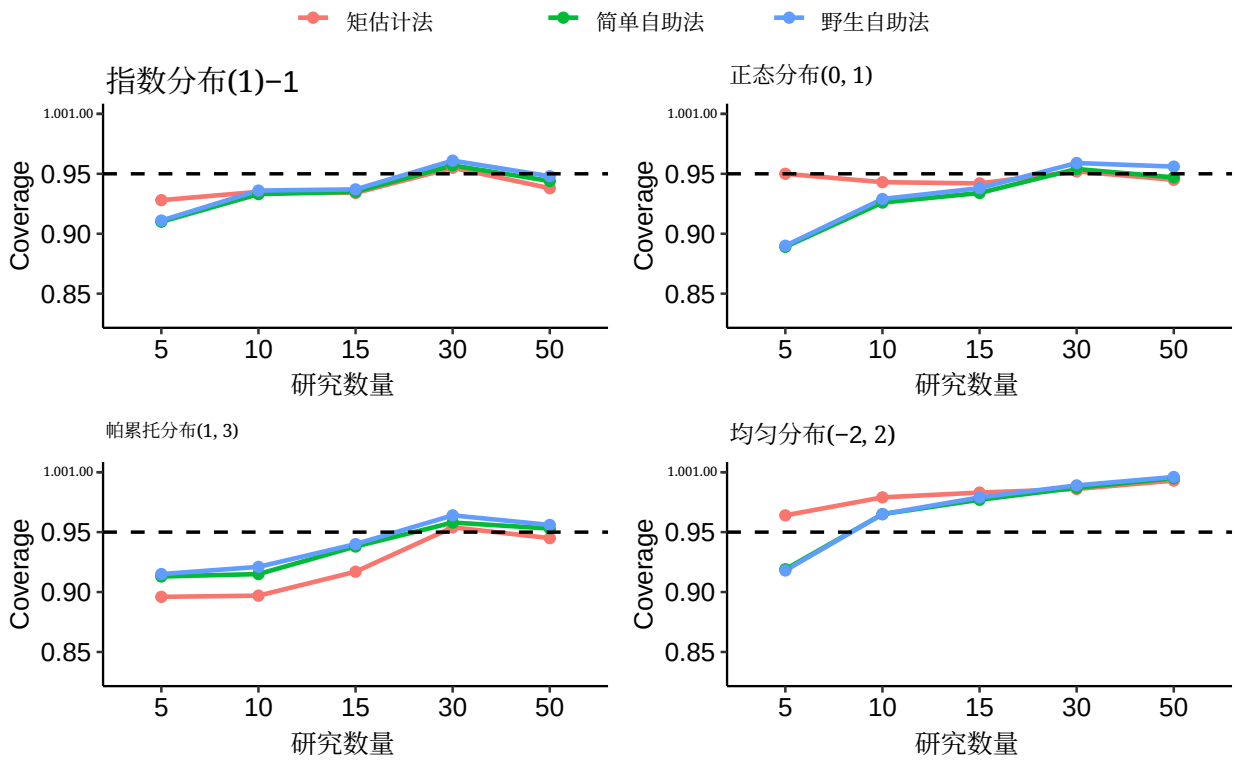


图2：根据正态分布的分位数构建的预测区间的覆盖率。每个图对应特定设置变异的一个单独分布。

5 讨论

在这项工作中，我们将因果可解释荟萃分析的最新进展加以扩展，以考虑超越试验与目标人群之间协变量差异的研究间异质性。我们的因果框架旨在弥合传统随机效应荟萃分析的结构与因果推断之间的差距。为此，我们提出了新的估计目标、估计值和推断程序，明确指出研究情境在造成研究特定参数的系统性差异方面可能发挥的作用。

研究间异质性的影响可能会严重限制将一组试验中的治疗效应估计值外推至目标人群时的可解释性。在存在此类异质性的情况下，基于不同试验合并数据得出的估计值，可能反映各试验情境中各种特有因素的综合影响。在许多情境下，即使将试验协变量分布标准化为目标人群的分布，这种异质性仍然存在。例如，即使在对所有效应修饰因子以及与治疗选择或试验参与相关的协变量进行条件化之后，治疗版本的差异仍可能导致我们预期在不同研究中观察到的潜在结局有所不同。

为解决这些问题，我们首先定义明确体现环境特定异质性的潜在结局，随后构建能够将这种额外变异纳入考量的估计量和推断程序。通过允许潜在结局同时取决于分配的治疗及该治疗分配所处的环境，我们可以将试验效应与试验所处人群的效应区分开来。

我们的工作是澄清因果可解释荟萃分析中试验环境与参与者双重影响的第一步。若能获得更多关于试验间方案层面差异的信息，我们的框架或许可以扩展，以纳入所关注的特定试验特征。未来的研究可以考虑迁移或提高某一特定异质性来源（例如依从性水平不同）的权重，从而得出与所关注临床环境最相关的迁移效应。与此相关，在试图将环境特定异质性纳入考量时，厘清意向治疗效应与符合方案效应之间的区别也很重要。也就是说，研究者应区分两种研究目标：考察目标人群成员参与特定试验且遵循方案时会发生什么，或考察其仅被分配接受治疗时会发生什么。本研究引入的估计目标应解释为界定意向治疗迁移效应，因为我们的观测数据仅包含治疗分配信息。不过，在有此类数据时，这些方法也可扩展用于描述符合方案效应。

我们的研究部分源于一个简单的认识：随机对照试验的具体情境与因果可解释性相关。换言之，从随机对照试验中迁移得出的估计值，其解释未必具有普遍性，也不能脱离试验情境的影响。我们认为，我们的框架是构建一种能够容纳研究间异质性的因果结构的第一步。未来，

我们计划进一步扩展这一结构，纳入与目标人群中各类临床问题相关的试验特定信息。

参考文献

Berlin, Jesse A. 和 Robert M. Golub。2014年。《作为证据的荟萃分析：构建更好的金字塔》。*JAMA* 312（6）：603。https://doi.org/10.1001/jama.2014.8167。Dahabreh, Issa J.、Lucia C. Petito、Sarah E. Robertson、Miguel A. Hernán 和 Jon A. Steingrimsson。2020年。《迈向因果可解释荟萃分析：将多项随机试验的推断迁移至新的目标人群》。*Epidemiology* 31（3）：334–44。https://doi.org/10.1097/EDE.0000000000001177。Dahabreh, Issa J.、Sarah E. Robertson 和 Miguel A. Hernán。2022。《对受不依从影响的治疗分配效应进行推广与迁移》。https://arxiv.org/abs/2211.04876。Dahabreh, Issa J.、Sarah E. Robertson、Lucia C. Petito、Miguel A. Hernán 和 Jon A. Steingrimsson。2022。《因果可解释荟萃分析的高效稳健方法：将多项随机试验的推断迁移至目标人群》。*Biometrics*。https://doi.org/https://doi.org/10.111 1/biom.13716。Dahabreh, Issa J.、James M. Robins、Sebastien J.-P. A. Haneuse 和 Miguel A. Hernán。2019。《推广随机试验中的因果推断：反事实识别与图形识别》。arXiv。http: //arxiv.org/abs/1906.10792。Degtiar, Irina 和 Sherri Rose。2021。《可推广性与可迁移性综述》。https: //doi.org/10.48550/ARXIV.2102.11904。DerSimonian, Rebecca 和 Nan Laird。1986年。《临床试验中的荟萃分析》。*Controlled Clinical Trials* 7（3）：177–88。Dhruva, Sanket S. 2008年。《临床试验参与者与美国医疗保险受益人在美国医疗保险国家覆盖决策所用证据方面的差异》。*Archives of Internal Medicine* 168（2）：136。https://doi.org/10.1001/archinternmed.2007.56。Ebell, Mark H、Jay Siwek、Barry D Weiss、Steven H Woolf、Jeffrey Susman、Bernard Ewigman 和 Marjorie Bowman。2004。《推荐强度分类法（SORT）：医学文献证据分级的以患者为中心的方法》。*The Journal of the American Board of Family Practice* 17（1）：59–67。Galson, Steven 和 Gregory Simon。2016。《指导新疗法审批与应用的真实世界证据》。*NAM Perspectives* 6（10）。https://doi.org/10.31478/201610b。

Higgins, Julian P. T.、Simon G. Thompson 和 David J. Spiegelhalter。2009。《重新评估随机效应荟萃分析》。皇家统计学会杂志。*A*辑（社会统计学） 172（1）： 137–59。http://www.jstor.org/stable/30136745。Kay, Stanley R.、Abraham Fiszbein 和 Lewis A. Opler。1987。《精神分裂症的阳性与阴性症状量表（PANSS）》。《精神分裂症通报》 13（2）： 261–76。https://doi.org/10.1093/schbul/13.2.261。Lunny, Carole、Cynthia Ramasubbu、Lorri Puil、Tracy Liu、Savannah Gerrish、Douglas M. Salzwedel、Barbara Mintzes 和 James M. Wright。2021年。《超过半数的临床实践指南使用非系统性方法为建议提供依据：一项方法学研究》。Tim Mathes 编。*PLOS ONE* 16（4）： e0250356。https://doi.org/10.1371/journal.pone.0250356。Matsouaka, Roland A.、Yi Liu 和 Yunji Zhou。2022。《治疗组和对照组平均治疗效应的方差估计》。arXiv。https://doi.org/10.48550/ARXIV.2209.10742。

Rao, Poduri S. R. S., Jack Kaplan 和 William G. Cochran。1981年。“误差方差不等时单因素随机效应模型的估计量。”美国统计学会杂志 76 (373): 89–97。http://www.jstor.org/stable/2287050。

Stadler, Eva, Khai Li Chai, Timothy E Schlub, Deborah Cromer, Mark N Polizzotto, Stephen J Kent, Nicole Skoetz 等。2022。“SARS-CoV-2感染中被动抗体有效性的决定因素。”*medRxiv*。https://doi.org/10.1101/2022.03.21.22272672。

Tsiatis, Anastasios A. 2006。“半参数理论与缺失数据。”美国食品药品监督管理局。2021年。“COVID-19公共卫生紧急事件期间医疗产品临床试验实施的FDA指南。”面向行业、研究者和机构审查委员会的指南 FDA-2020-D-1106-0002。Vale, C. L., L. H. M. Rydzewska, M. M. Rovers, J. R. Emberson, F. Gueyffier, L. A. Stewart, 以及代表Cochrane个体参与者数据荟萃分析方法组。2015。“临床实践指南对基于个体参与者数据的系统综述和荟萃分析的采用情况：描述性研究。”《英国医学杂志》 350 (mar06 6): h1088–88。https://doi.org/10.1136/bmj.h1088。VanderWeele, Tyler J., 和 Miguel A. Hernán。2013。“多种治疗版本下的因果推断。”《因果推断杂志》 1 (1): 1–20。https://doi.org/doi:10.1515/jci-2012-0002。

致谢

作者感谢明尼苏达大学的明尼苏达超级计算研究所（MSI）提供资源，为本文报告的研究成果作出了贡献。网址：http: //www.msi.umn.edu

6 附录

6.1 识别公式结果的完整证明，适用于$\mu_{a,\ 0}(k_s^a)$

$$
\begin{aligned}
E[Y(a,k_s^a)|R=0] &= E[E[Y(a,k_s^a)|X,R=0]|R=0] && \text{Law of total expectation}\\
&= E[E[Y(a,k_s^a)|X,S=s]|R=0] && \text{By Assumption 1}\\
&= E[E[Y(a,k_s^a)|X,S=s,A=a]|R=0] && \text{By Assumption 2}\\
&= E[E[Y(a,k_s^a)|X,S=s,A=a,K^a(a)=k_s^a]|R=0] && \text{By Assumption 5}\\
&= E[E[Y|X,S=s,A=a]|R=0] && \text{By Assumption 3}
\end{aligned}
$$

6.2 $E[Y（a，\ K_0^a（a））|R=0]$的近似的完整推导

$$
\begin{aligned}
E[Y(a,K_0^a(a))|R=0] &= \sum_{k^a\in\mathcal{K}^a} E\left\{E[Y(a,K_0^a(a))|R=0,K_0^a(a)=k^a]|R=0\right\}P(K_0^a(a)=k^a)\\
&\approx \sum_{k^a\in\mathcal{K}^a} E\left\{E[Y(a,K_0^a(a))|R=0,K_0^a(a)=k^a]|R=0\right\}\hat{P}(K_0^a(a)=k^a)\\
&= \sum_{k^a\in\mathcal{K}^a} E\left\{E[Y(a,K_0^a(a))|R=0,K_0^a(a)=k^a]|R=0\right\}\left\{\frac{1}{m}\sum_{s=1}^{m}\mathbb{1}(k^a=k_s^a)\right\}\\
&= \sum_{k^a\in\mathcal{K}^a}\frac{1}{m}\sum_{s=1}^{m}\mathbb{1}(k^a=k_s^a)E\left\{E[Y(a,K_s^a(a))|R=0,K_s^a(a)=k^a]|R=0\right\}\\
&= \frac{1}{m}\sum_{s=1}^{m}\sum_{k^a\in\mathcal{K}^a}\mathbb{1}(k^a=k_s^a)E\left\{E[Y(a,K_s^a(a))|R=0,K_s^a(a)=k^a]|R=0\right\}\\
&= \frac{1}{m}\sum_{s=1}^{m}E\left\{E[Y(a,K_s^a(a))|R=0,K_s^a(a)=k_s^a]|R=0\right\}\\
&= \frac{1}{m}\sum_{s=1}^{m}E[Y(a,k_s^a)|R=0]\\
&= \frac{1}{m}\sum_{s=1}^{m}E[E[Y|X,S=s,A=a]|R=0] && \text{By (7)}
\end{aligned}
$$

6.3 存在……时我们对研究间方差的估计值的完整推导

相关估计量

采用主稿中定义的符号，将我们估计量的总平方和写为

$$
Q=\sum_{s=1}^{m}(\hat{\mu}_s-\hat{\mu})^2=\sum_{s=1}^{m}\left(\hat{\mu}_s-\sum_{l=1}^{m}w_l\hat{\mu}_l\right)^2
$$

and

$$Q_s = (\hat{\mu}_s - \hat{\mu})^2 = \left(\hat{\mu}_s - \sum_{l=1}^{m} w_l \hat{\mu}_l\right)^2 .$$

因此， $Q = \sum_{s=1}^{m} Q_s$。与 Rao、Kaplan 和 Cochran（1981年）一样，我们推导出Q 的期望表达式以及 γ^2 的估计量。为此，令

$$\mathbf{a}_s = \begin{pmatrix} -w_1 \\ \vdots \\ 1 - w_s \\ \vdots \\ -w_m \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \sigma_1^2 + \gamma^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 + \gamma^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 + \gamma^2 \end{pmatrix}$$

为 $\hat{\boldsymbol{\mu}}$ 的方差-协方差矩阵，其中 $\sigma_i^2 + \gamma^2$ 是第 i 个对角元素，σ_{ij} 是第 (i, j) 个非对角元素（即 σ_{ij} 是 $\mathrm{Cov}(\hat{\mu}_i, \hat{\mu}_j)$）。然后，

$$\begin{aligned} \mathbf{a}_s \mathbf{a}_s^T \Sigma &= \begin{pmatrix} -w_1 \\ \vdots \\ 1 - w_s \\ \vdots \\ -w_m \end{pmatrix} \begin{pmatrix} -w_1 & \cdots & 1 - w_s & \cdots - w_m \end{pmatrix} \begin{pmatrix} \sigma_1^2 + \gamma^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 + \gamma^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 + \gamma^2 \end{pmatrix} \\ &= \begin{pmatrix} w_1^2 & \cdots & -w_1(1 - w_s) & \cdots & w_1 w_m \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -w_1(1 - w_s) & \cdots & (1 - w_s)^2 & \cdots & -w_m(1 - w_s) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_1 w_m & \cdots & -w_m(1 - w_s) & \cdots & w_m^2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 + \gamma^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 + \gamma^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 + \gamma^2 \end{pmatrix} . \end{aligned}$$

也就是说，计算出$\mathbf{a}_s \mathbf{a}_s^T$后，我们得到一个矩阵，其中第 i 个对角线元素为 w_i^2，但第 s 个对角线元素除外，它为 $(1 - w_s)^2$。对于每一行 $i, i \neq s$，元素 (i, j)、$j \neq s$ 和 $j \neq i$ 均由 $w_i w_j$ 给出，而元素 (i, s) 为 $-w_i(1 - w_s)$。在第 s 行中，每个元素 (s, j) $j \neq s$ 均为 $-w_j(1 - w_s)$。要计算 $E[Q_s]$，我们需要

计算 $\mathbf{tr}$（$\mathbf{a}_s\mathbf{a}_s^T\Sigma$）。观察上面第二个等式中的矩阵，对权重矩阵的第一行与 Σ 的第一列取点积，可以看出第 i 个这样的对角线元素（$i \neq s$）

$$w_i^2(\sigma_i^2+\gamma^2) - w_i(1-w_s)(\sigma_{is}) + \sum_{j\neq i, j\neq s} w_i w_j \sigma_{ij}.$$

第 s 个对角元素为

$$(1-w_s)^2(\sigma_s^2+\gamma^2) + \sum_{i\neq s}(-w_i)(1-w_s)\sigma_{is}.$$

将所有对角元素相加,

$$\begin{aligned}\mathrm{tr}(\mathbf{a}_s\mathbf{a}_s^T\Sigma) &= (1-w_s)^2(\sigma_s^2+\gamma^2) + \sum_{i\neq s}(-w_i)(1-w_s)\sigma_{is} + \sum_{i\neq s}\left[w_i^2(\sigma_i^2+\gamma^2) - w_i(1-w_s)(\sigma_{is}) + \sum_{j\neq i, j\neq s} w_i w_j\sigma_{ij}\right] \\ &= (1-2w_s)(\sigma_s^2+\gamma^2) + \sum_{i=1}^m w_i^2(\sigma_i^2+\gamma^2) - 2\sum_{i\neq s} w_i(1-w_s)\sigma_{is} + \sum_{i\neq s}\left[\sum_{j\neq i, j\neq s} w_i w_j\sigma_{ij}\right].\end{aligned}$$

No那么问题是：如何利用矩方程计算 γ^2？ 首先，令

$$C_s = -2\sum_{i\neq s} w_i(1-w_s)\sigma_{is} + \sum_{i\neq s}\left[\sum_{j\neq i, j\neq s} w_i w_j\sigma_{ij}\right]$$

表示 $\mathbf{tr}(\mathbf{a}_s\mathbf{a}_s^T\Sigma)$ 中所有关于 γ^2 为常数的分量。然后，注意到$\boldsymbol{\mu}^T\mathbf{a}_i\mathbf{a}_i^T\boldsymbol{\mu} = 0$，我们有

$$E[Q_s] = (1-2w_s)(\sigma_s^2+\gamma^2) + \sum_{i=1}^m w_i^2(\sigma_i^2+\gamma^2) + C_s$$

而且,

$$\begin{aligned}E[Q] &= \sum_{s=1}^m E[Q_s] \\ &= \sum_{s=1}^m\left[(1-2w_s)(\sigma_s^2+\gamma^2) + \sum_{i=1}^m w_i^2(\sigma_i^2+\gamma^2) + C_s\right] \\ &= m\sum_{s=1}^m w_s^2(\sigma_s^2+\gamma^2) + \sum_{s=1}^m\left[(1-2w_s)(\sigma_s^2+\gamma^2) + C_s\right].\end{aligned}$$

我们将此表达式用于 $E[Q]$ ，以推导论文正文中给出的研究间异质性估计值。

6.4 逆概率加权估计量的识别

这一识别结果的推导与 Dahabreh 等人（2022）中类似的逆概率加权识别结果相似。重要的区别在于，每项试验使用的权重各不相同，而不是将权重应用于试验人群的所有成员。下面的结果由迭代期望和我们的正性假设（假设4）推得。

$$
\begin{aligned}
\psi_{s,0}(a) &= E[E[Y|X, S=s, A=a]|R=0] \\
&= E\left[E\left[\frac{I(S=s, A=a)Y}{P(S=s|X)P(A=a|X, S=s)}\middle| X\right]\middle| R=0\right] \\
&= \frac{1}{P(R=0)}E\left[I(R=0)E\left[\frac{I(S=s, A=a)Y}{P(S=s|X)P(A=a|X, S=s)}\middle| X\right]\right] \\
&= \frac{1}{P(R=0)}E\left[E\left[\frac{I(S=s, A=a)YP(R=0|X)}{P(S=s|X)P(A=a|X, S=s)}\middle| X\right]\right] \\
&= \frac{1}{P(R=0)}E\left[E\frac{I(S=s, A=a)YP(R=0|X)}{P(S=s|X)P(A=a|X, S=s)}\right]
\end{aligned}
$$

6.5 完整模拟结果：基于分位数的预测区间

		矩估计法		简单自助法		野生自助法	
N 研究	.....的分布 德尔塔	覆盖率	长度	覆盖率	长度	覆盖率	长度
5	指数分布(1)-1	92.8%	5.55	88.9%	3.24	88.8%	3.29
5	正态分布(0, 1)	95%	5.91	83.5%	3.45	83.8%	3.50
5	帕累托分布(1, 3)	89.6%	3.84	90.9%	2.50	91%	2.54
5	均匀分布(-2, 2)	96.4%	7.03	86.6%	3.91	87.5%	3.95
10	指数分布(1)-1	93.5%	4.38	93.6%	3.68	94.4%	3.77
10	正态分布(0, 1)	94.3%	4.51	89.4%	3.79	91%	3.87
10	帕累托分布(1, 3)	89.7%	2.80	92.7%	2.65	93.5%	2.76
10	均匀分布(-2, 2)	97.9%	5.26	93.8%	4.03	94.2%	4.12
15	指数分布(1)-1	93.4%	4.08	95%	3.83	95.5%	3.95
15	正态分布(0, 1)	94.2%	4.30	92%	4.00	92.7%	4.08
15	帕累托分布(1, 3)	91.7%	2.74	95%	2.90	95.5%	3.02
15	均匀分布(-2, 2)	98.3%	4.98	94.8%	4.07	95.7%	4.18
30	指数分布(1)-1	95.5%	3.96	98.1%	4.20	97.7%	4.26
30	正态分布(0, 1)	95.2%	4.07	94.2%	4.16	94.5%	4.19
30	帕累托分布(1, 3)	95.4%	2.81	97.5%	3.46	97.6%	3.54
30	均匀分布(-2, 2)	98.6%	4.71	96.2%	4.03	98.2%	4.18
50	指数分布(1)-1	93.8%	3.96	96.9%	3.96	96.8%	4.13
50	正态分布(0, 1)	94.5%	4.00	94%	4.06	94.6%	4.18
50	帕累托分布(1, 3)	94.5%	2.98	96%	2.97	96.7%	3.21
50	均匀分布(-2, 2)	99.3%	4.64	97.6%	4.03	98.8%	4.21

6.6 完整模拟结果：基于正态分布的预测区间 s

		矩估计法		简单自助法		野生自助法	
N 研究	分布 德尔塔	覆盖率	长度	覆盖率	长度	覆盖率	长度
5	指数分布(1)-1	92.8%	5.55	91%	4.04	91.1%	4.06
5	正态分布(0, 1)	95%	5.91	88.9%	4.25	89%	4.26
5	帕累托分布(1, 3)	89.6%	3.84	91.3%	3.03	91.5%	3.03
5	均匀分布(-2, 2)	96.4%	7.03	91.9%	4.99	91.8%	5.00
10	指数分布(1)-1	93.5%	4.38	93.3%	4.09	93.6%	4.13
10	正态分布(0, 1)	94.3%	4.51	92.6%	4.19	92.9%	4.23
10	帕累托分布(1, 3)	89.7%	2.80	91.5%	2.81	92.1%	2.88
10	均匀分布(-2, 2)	97.9%	5.26	96.5%	4.83	96.5%	4.86
15	指数分布(1)-1	93.4%	4.08	93.5%	3.99	93.7%	4.05
15	正态分布(0, 1)	94.2%	4.30	93.4%	4.17	93.8%	4.23
15	帕累托分布(1, 3)	91.7%	2.74	93.8%	2.83	94%	2.92
15	均匀分布(-2, 2)	98.3%	4.98	97.7%	4.79	97.9%	4.85
30	指数分布(1)-1	95.5%	3.96	95.7%	4.00	96.1%	4.08
30	正态分布(0, 1)	95.2%	4.07	95.4%	4.09	95.9%	4.17
30	帕累托分布(1, 3)	95.4%	2.81	95.8%	2.95	96.4%	3.08
30	均匀分布(-2, 2)	98.6%	4.71	98.7%	4.69	98.9%	4.77
50	指数分布(1)-1	93.8%	3.96	94.4%	4.03	94.8%	4.14
50	正态分布(0, 1)	94.5%	4.00	94.7%	4.07	95.6%	4.17
50	帕累托分布(1, 3)	94.5%	2.98	95.3%	3.12	95.6%	3.29
50	均匀分布(-2, 2)	99.3%	4.64	99.5%	4.68	99.6%	4.78