· 综述 ·

结合倾向性评分的贝叶斯外对照借用方法学研究进展*

王　锴[1]　曹　寒[2]　刘天谋[1,3]　于永沛[1]　阎小妍[1]　姚　晨[1,2△]

【提　要】 随机对照临床试验(randomized clinical trial,RCT)是验证药物和医疗器械有效性和安全性的金标准,但对于罕见病、肿瘤以及儿科人群等特殊领域,开展 RCT 研究往往面临着样本量不足、研究周期过长以及医学伦理方面的阻碍。近年来,外部对照试验作为 RCT 的替代方案得到了越来越多的关注。开展外部对照试验需要着重解决两个关键问题:①如何确定外部对照组的信息借用程度;②研究间受试者基线协变量的可比性。贝叶斯统计分析方法在外部对照试验中得到了广泛的应用,但其只考虑了内外部对照组边际结局的实际分布差异,并未将基线协变量的可比性纳入其分析框架,从而容易引入选择偏倚。为了解决上述问题,一些学者提出了结合倾向性评分的贝叶斯外对照借用方法,即在使用贝叶斯方法前先通过倾向性评分匹配、分层和加权等手段构造基线协变量可比的伪人群,以达到控制选择偏倚的目的。本文将从倾向性评分方法学入手,对目前"结合倾向性评分的贝叶斯外对照借用方法"进行介绍,包括其基本原理、分析步骤以及统计学操作特性等,旨在为后续外部对照试验的开展提供统计设计和分析的方法学支持。

【关键词】 外部对照试验　倾向性评分　贝叶斯统计　动态借用　选择偏倚　I 类错误
【中图分类号】 R195.1　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.03.030

在药械研发过程中,需要通过严格设计的临床试验对其有效性和安全性进行充分评价,其中随机对照临床试验(randomized clinical trial,RCT)是金标准[1]。但对于罕见病、肿瘤以及儿科人群等特殊领域,开展 RCT 研究往往面临着样本量不足、研究周期过长以及医学伦理方面的阻碍[2-6]。因此,越来越多的学者开始探讨如何利用现有的外部临床研究数据来支持 RCT 研究的统计设计与分析,其中外部对照试验(externally controlled trials)是当前的研究热点之一[7]。

与平行对照不同,外部对照组(external control,EC)并非属于试验组所在的随机试验,而是来自于本研究以外的一组患者[8-9],他们可以是早些时候接受过治疗的一组患者(非同期外部对照组或历史对照)或是在同一时间不同环境下接受治疗的一组患者[10]。在 ICH E10 中,外部对照试验指将接受试验治疗的一组对象与本研究以外的一组患者而不是与分配到不同治疗组的相同人群患者组成的内部对照组进行比较,因此也被称作"完全增强设计"或单臂试验[7-9]。除此之外,外部增强 RCT 设计是近年来提出的另一种外部对照试验设计类型。其是指对照组中既有新纳入的,也有来自现有外部临床数据的受试者[7],因此也被称作混合对照臂(hybrid control arm)。此外还有混合 RCT 设计,其在外部增强 RCT 设计之上重点强调了灵活性,常与期中分析相结合,当外部对照组与当前研究高度匹配时才决定使用,否则继续开展传统的 RCT 研究[7]。这两种设计均在不平衡 RCT 设计的基础上,通过借用外部对照组数据来实现内部对照组样本量扩充的目的。真实世界数据(real world data,RWD)也是外部对照组的重要来源之一,例如电子病例、疾病登记数据库和保险索赔数据库等[5]。

开展外部对照试验需要着重解决两个关键问题,首先是确定外部对照组的信息借用程度[11]。鉴于贝叶斯方法所具有的灵活性以及其自身含有的数据更新思想,其特别适合于外部对照试验的统计分析。使用贝叶斯外对照借用方法的前提是满足可交换性假设[12]。当内外部对照组间边际结局的实际分布不一致时,即出现"先验—数据矛盾"(prior-data conflict),可交换性假设将受到挑战,此时要减少从外部对照组借用的信息量[13]。目前,大多数贝叶斯外对照借用方法根据内外部对照组间边际结局的实际分布差异程度决定外部对照组的信息借用量。这种基于数据驱动的客观指导外部对照组信息借用量的方法也被称作是动态借用(dynamic borrowing)[13],例如改良幂先验法(modified power prior,MPP)[14]、相称性先验(commesurate prior,CP)[15]、校正先验(calibrate prior)[16]以及 meta 分析预测先验(meta-analytical predict prior,MAP)[17-18]等。

开展外部对照试验还需要重点关注当前研究与外部对照组受试者在基线协变量上的可比性[11]。ICH E10 指出,保证受试者基线特征可比是选择外部对照组需要满足的条件之一,有利于减少因缺少随机化而带来的选择偏倚[8-9]。然而单纯使用贝叶斯外对照借

* 基金项目:海南省博鳌乐城先行区真实世界研究专项计划项目(HNLC2022RWS017)
1. 北京大学第一医院北京大学临床研究所(100191)
2. 北京大学第一医院医学统计室
3. 北京大学公共卫生学院
△通信作者:姚晨,E-mail:yaochen@hsc.pku.edu.cn

用方法并未涉及对基线协变量可比性的考察，从而容易引入选择偏倚。倾向性评分方法学是当前处理基线协变量分布不可比问题的常用工具。因此一些学者顺势提出了“结合倾向性评分的贝叶斯外对照借用方法”。本文首先将在简要回顾倾向性评分方法学的基础上，对目前主要的“结合倾向性评分的贝叶斯外对照借用方法”进行介绍，包括其基本原理、分析步骤以及统计学操作特性。

倾向性评分方法学

倾向性评分（propensity score，PS）最初由 Rosenbaum 和 Rubin 提出[19]，对于双臂研究设计而言，其指的是具有某种基线协变量特征 X 的受试者被分配到试验组（$W=1$）的条件概率，即：

$$e(X)=\Pr(W=1|X) \quad (1)$$

基于强可忽视性假设，PS 具有平衡得分的性质，对于 PS 相同的受试者，其基线协变量分布理论上是一致的[19]。通常可以利用 logistic 回归模型对倾向性评分进行估计，其中基线协变量为自变量，是否分配到试验组（$W=1$）为因变量。倾向性评分在观察性研究中的应用十分广泛，常见的倾向性评分方法学包括利用 PS 进行匹配、分层、加权以及直接将 PS 作为协变量放入模型等[20]。

对于外部对照试验而言，倾向性评分的含义有所改变，其指的是具有某种基线协变量特征 X 的受试者被纳入到当前研究（$S=1$）的条件概率，即：

$$e^*(X)=\Pr(S=1|X) \quad (2)$$

有些研究也将其称作是“在试得分”（on-trial score）[21]。同样，具有相同倾向性评分 $e^*(X)$ 的受试者，其基线协变量的分布理论上是一致的。

结合倾向性评分的贝叶斯外对照借用方法学

为了符合监管机构和相关指南的要求，尽可能地模拟 RCT 设计并满足“前瞻性设计”的原则[22-24]。目前倾向性评分方法学与贝叶斯外对照借用方法学的结合策略为“序贯两阶段法”，即分为“设计”阶段和“分析”阶段，目的是将基线协变量信息与结局信息区分开来[11, 21, 24]。其中，“设计”阶段主要包含倾向性评分的估计以及通过匹配、分层以及加权的方法构建协变量可比的伪人群，此时不涉及任何结局信息；在“分析”阶段，主要涉及贝叶斯外对照借用方法的使用，如先验分布形式的确定[11, 21, 24]。下文记 D_0 为外部对照组、D_1 为当前研究，参数 θ 为目标参数。

1. 倾向性评分匹配（PS matching）

PS 匹配法将从外部对照组中选取倾向性评分相近的受试者，具体包括贪婪匹配（greedy matching）和最优匹配（optimal matching）两种策略。前者要求每一个待匹配受试者与匹配得到的受试者间 PS 距离是最小的，后者则要求待匹配受试者与匹配得到的受试者的组间 PS 整体距离是最小的[25-26]。通常外部对照组的样本量较大（如真实世界研究数据），因此使用最优匹配的效果往往比不上贪婪匹配法[27]。

（1）幂先验

Lin 等人[25]将 PS 匹配与幂先验法相结合，并应用于外部增强 RCT 设计。在“设计”阶段，Lin 通过非参数模型估计倾向性评分并进行 PS 匹配，从而构造与当前研究试验组可比的对照组，模拟 1∶1 的 RCT 设计。Lin 等人提出了两种匹配策略，分别为基于最优匹配的配对匹配法以及基于贪婪匹配的最近邻卡钳值匹配法。在“分析”阶段，Lin 等人直接使用匹配后外部对照组受试者的倾向性评分估计值作为幂先验中似然函数的折扣系数，于是目标参数 θ 的幂先验分布的表达式为：

$$p(\theta|D_0,\delta) \propto \prod_{j=1}^{N_{CH}} L(\theta|D_0)_j^{\delta}\pi(\theta) \quad (3)$$

其中，N_{CH} 为匹配后得到的外部对照组样本量，参数 δ_j 为折扣系数，且 $\delta_j=e^*(X_j)$，$\pi(\theta)$ 为参数的初始先验。

Harton 等人[27]指出了该方法可能存在以下缺点：首先，对于来自 RWD 的外部对照组，因其样本量较大，使用倾向性评分匹配可能遇到困难；其次，在匹配的基础上又利用倾向性评分作为折扣系数，进一步减少最终借用的样本量，无法真正模拟 1∶1 的 RCT 设计；最后，直接使用倾向性评分作为折扣系数相当于对匹配后对照组进行加权，然而实际并没有与之相对应的因果效应估计量。

（2）相称性先验

Wang 等人[21]将 PS 匹配与相称性先验法相结合，并应用于外部增强 RCT 设计。在“设计”阶段，其沿用了 Lin 等人提出的 1∶1 最近邻卡钳值匹配策略。在“分析”阶段，以二分类结局变量 Y 为例，先对目标参数进行 logit 变换，并假设转换后的参数服从正态分布，于是目标参数的相称性先验分布的表达式为：

$$p(\mathrm{logit}(\theta_{C_0}),\mathrm{logit}(\theta_h),\tau|D_0) \propto L(\theta_h|D_0)N(\mathrm{logit}(\theta_{c_0})|\mathrm{logit}(\theta_h),\tau)\pi(\mathrm{logit}(\theta_h))\pi(\tau) \quad (4)$$

其中，参数 τ 为衡量相似性大小的参数且 $\pi(\tau)$ 为其初始先验；参数 θ_{C_0} 和 θ_h 分别代表内部对照组以及匹配后外部对照组的结局变量分布参数，其中参数 θ_{C_0} 为目标参数；似然函数 $L(\theta_h|D_0)=\prod_{j\in N_{CH}}\{\theta_h^{Y_j}(1-\theta_h)^{1-Y_j}\}$，其中 N_{CH} 为经匹配得到的外部对照组样本量；$N(\mathrm{logit}(\theta_{c_0})|\mathrm{logit}(\theta_h),\tau)$ 为均值为 $\mathrm{logit}(\theta_h)$，方差为 τ 的正态分布。

2. 倾向性评分分层(PS stratification)

与 PS 匹配法不同,PS 分层法不会剔除外部对照组中的受试者,而是将其按照倾向性评分的大小进行分层,从而保证层内受试者的基线协变量分布的一致性。

(1)改良幂先验

Wang 等人[28]将 PS 分层与改良幂先验法相结合,并将其应用于单臂(仅试验组)研究借用外部 RWD(试验组)进行样本量扩充的场景。

在“设计”阶段,Wang 等人根据 PS 对当前研究受试者进行分层,假设共分为 K 层。在经裁剪(trimming)后,将 PS 超出当前研究受试者倾向性评分范围之外的外部对照组受试者剔除。之后,Wang 等人引入参数v_k 作为该层内外部受试者基线协变量一致程度的度量指标,并要求其满足$v_k \geqslant 0$ 且 $\sum_{1}^{K} v_k = 1$。由于组间倾向性评分分布重合度越高,内外部受试者的基线协变量分布越一致,因此 Wang 等人将参数v_k 与第 k 层内外部受试者的倾向性评分分布重合度相结合。记第 k 层 PS 的经验分布函数和密度函数分别为$F_{k,s}$和$f_{k,s}$(其中 $S=1$ 代表当前研究),那么该层内外部受试者的倾向性评分分布重合度r_k 的计算公式为:

$$r_k = \int_0^1 \min[f_{k,0}(e(X)) - f_{k,1}(e(X))]de(X) \quad (5)$$

实际上,除了倾向性评分分布重合度外,还可以利用其他能够反应两个分布相似程度的指标,例如 Kolmogorov-Smirnov 距离等。

在“分析”阶段,Wang 等人通过层内相似程度v_k 定义改良幂先验的折扣系数 δ。记第 k 层的幂先验折扣系数为 δ_k:

$$\delta_k = \min(\frac{Av_k}{n_{k,h}}, 1) \quad (6)$$

其中,参数 A 与参数$n_{k,h}$分别代表计划从外部 RWD 中借用的总样本量和该层中外部 RWD 的受试者人数。基于第 k 层层内倾向性评分分布重合度r_k,有两种计算层内相似程度v_k 的方法,分别为固定系数法和全贝叶斯法。

当采用固定系数法计算层内相似程度v_k 时层内相似程度$v_k = \frac{r_k}{\sum_{k=1}^{K} r_k}$,目标参数的幂先验分布表达式为:

$$\pi(\theta_1, \cdots, \theta_k \mid D_0) \propto \prod_{k=1}^{K} [L(\theta_k \mid D_{k,0})^{\delta_k}] \pi(\theta_1, \cdots, \theta_k) \quad (7)$$

由于监管机构需要事先确定从外部数据中借用的样本量,因此固定系数法更加适合。但是,由于在“设计”阶段就确定了折扣系数,因此该方法不具备动态借用的性质。

当使用全贝叶斯法时,需要对参数v_k 指定初始先验分布,例如 Dirichlet 先验 $\pi(v_1, \cdots, v_k) = Dir(\frac{r_1}{R}, \cdots, \frac{r_k}{R})$,其中参数 R 控制了先验的方差大小,此时目标参数的改良幂先验分布表达式为:

$$\pi(\theta_1, \cdots, \theta_k, v_1, \cdots, v_k \mid D_0) \propto \frac{\prod_{k=1}^{K} [L(\theta_k \mid D_{k,0})^{\delta_k}] \pi(\theta_1, \cdots, \theta_k) \pi(v_1, \cdots, v_k)}{\int \prod_{k=1}^{K} [L(\theta_k \mid D_{k,0})^{\delta_k}] \pi(\theta_1, \cdots, \theta_k) d\theta_1, \cdots, d\theta_k} \quad (8)$$

此时每一层具体借用的受试者样本量由实际数据决定,因此具有动态借用的性质。此时初始先验 $\pi(v_1, \cdots, v_k)$实际上是一种弱信息先验分布,其分布参数依赖于r_k。该处理使得每一层内外部受试者间基线协变量越可比,该层折扣系数的初始先验均值将越大,于是偏向于借用更多的样本量。

基于上述两种方法可以分别得到目标参数的后验分布 $\pi(\theta_1, \cdots, \theta_k \mid D_0, D_1)$ 和 $\pi(\theta_1, \cdots, \theta_k, v_1, \cdots, v_k \mid D_0, D_1)$。最后通过对各层参数 $\theta_1, \cdots, \theta_k$ 分布进行加权平均后可以得到目标参数 θ 的分布。一般而言,将当前研究中该层受试者所占的比例$\frac{n_{c_0,k}}{n_{c_0}}$作为该层参数的权重。特别的,当采取等比例倾向性评分分层时,参数 θ 的加权平均分布为$\frac{\sum_{k=1}^{K} \theta_S}{K}$。

模拟研究结果显示,固定比例法与全贝叶斯法相比,前者的参数估计偏差较大但均方误差更小。当协变量的个数变多时,无论采用何种方法确定折扣系数,参数估计的偏差与均方误差均增大。此外,Lu 等人[11]在此基础上还进一步提出了针对多个外部对照组的统计分析框架。

(2)贝叶斯层次先验

Baron 等人[29]延续 Wang 等人的思路,提出了“基于倾向性评分的贝叶斯逐个击破”分析框架(bayesian divide - and - conquer propensity score based approaches)。该分析框架含有两个步骤,分别是“倾向性评分分层(divide 步)”和“分析(analysis 步)”,且同样可应用于单臂(仅试验组)研究借用外部 RWD(试验组)进行样本量扩充的场景。其中,divide 步的方法同 Wang 等人的方法一致。在 analysis 步,Baron 等人利用了贝叶斯层次先验法。

对于第 k 层,假设连续型结局变量满足$\bar{Y}_{k,s} \mid \theta_{k,s} \sim N(\theta_{k,s}, \frac{\sigma_{k,s}^2}{n_{k,s}})$、$\frac{(n_{k,s}-1)s_{k,s}^2}{\sigma_{k,s}^2} \sim \chi^2_{n_{k,s}-1}$,其中 $s=1$ 代表当前

研究，$s=0$ 代表外部 RWD，且 $\bar{Y}_{k,s}$ 和 $s^2_{k,s}$ 分别表示该层内结局变量的样本均值和样本方差。Baron 等人针对目标参数 $\theta_{k,s}$ 设计了贝叶斯层次先验，其中第一层先验为：

$$\theta_{k,s}\sim N(\mu_k,\tau_k^2),s=0,1 \tag{9}$$

第二层先验为：

$$\begin{cases}\mu_k\sim N(\lambda_k,\varphi_k^2)\\ \lambda_k|\varphi_k^2\sim N(0,\kappa_0\varphi_k^2)\\ \varphi_k^2\sim \text{iGamma}(a,b)\\ \tau_k^2\sim \text{truncated normal}(b_1,b_2,b_3,b_4)\end{cases} \tag{10}$$

其中，层内异质性参数 τ_k^2 的初始先验选择为截断正态分布。基于上述贝叶斯层次先验，记 $D_s=(\bar{Y}_{k,s},s^2_{k,s})$，参数 $(\theta_{k,1},\theta_{k,0},\lambda_k,\varphi_k^2,\tau_k^2,\sigma^2_{k,0},\sigma^2_{k,1})=\Lambda$，于是第 k 层后验分布 $\pi(\Lambda|D_0,D_1)$ 的表达式为：

$$\begin{gathered}\pi(\Lambda|D_0,D_1)\propto L(\theta_{k,1},\sigma^2_{k,1}|D_1)L(\theta_{k,0},\sigma^2_{k,0}|D_0)\\ \pi(\theta_{k,0}|\mu_k,\tau_k^2)\pi(\theta_{k,1}|\mu_k,\tau_k^2)\cdot\\ \pi(\mu_k|\lambda_k,\varphi_k^2)\pi(\lambda_k,\varphi_k^2)\pi(\tau_k^2)\pi(\sigma^2_{k,0})\pi(\sigma^2_{k,1})\end{gathered} \tag{11}$$

除层次先验外，Liu 等人[30]以及 Zhu 等人[31]还将 PS 分层与 MAP 先验相结合，该方法适用于存在多个外部对照组的应用场景，具体请参考相应文献。

(3)相称性先验

Wang 等人[21]在外部增强 RCT 设计下，还将相称性先验与 PS 分层相结合。依旧以二分类结局变量 Y 为例，在“分析”阶段需要先对目标参数进行 logit 变换，并假设转换后的参数服从正态分布，于是可得第 k 层目标参数的相称性先验表达式：

$$\begin{gathered}p(\text{logit}(\theta_{c_0,k}),\text{logit}(\theta_{h,k}),\tau_k|D_{0,k})\propto L(\theta_{h,k}|D_0)\\ N(\text{logit}(\theta_{c_0,k})|\text{logit}(\theta_{h,k}),\tau_k)\times\pi(\text{logit}(\theta_{h,k}))\pi(\tau_k)\end{gathered} \tag{12}$$

其中，参数 τ_k 反映各层内外部对照组间异质性的大小且 $\pi(\tau_k)$ 为其初始先验，参数 $\theta_{c_0,k}$ 以及 $\theta_{h,k}$ 分别为第 k 层内外部对照组结局变量的分布参数；似然函数 $L(\theta_{h,k}|D_0)=\prod_{j\in N_k}\{\theta_{h,k}{}^{Y_j}(1-\theta_{h,k})^{1-Y_j}\}$，其中 N_k 为第 k 层外部对照组样本量；$N(\text{logit}(\theta_{c_0,k})|\text{logit}(\theta_{h,k}),\tau_k)$ 为均值为 $\text{logit}(\theta_{h,k})$，方差为 τ_k 的正态分布。

3.倾向性评分加权(PS weighting)

PS 加权同样不会剔除外部对照组的受试者，而是通过赋予一定的权重，使得加权后的外部对照组受试者的基线协变量分布与内部对照组一致。常用的倾向性评分加权包括逆概率加权法(inverse probability weighting，IPW)和标准化死亡率加权(standardized mortality ratio weighting，SMRW)[32]。前者对于内外部对照组中所有的受试者进行加权，得到的因果效应为“平均处理效应(average treatment effect，ATE)”；相反，后者只对外部对照组受试者进行加权，而内部对照组的权重为 1，此时得到的因果效应为“试验组人群的平均处理效应”(average treatment effect on treated，ATT)[32]。

(1)幂先验

① 标准化死亡率加权法

Wang 等人[21]在外部增强 RCT 设计下，将倾向性评分 SMRW 与幂先验法相结合。

在“设计”阶段，Wang 等人采用标准化 SMRW 权重构建伪人群。其中，内部对照组的权重均为 1，而外部对照组的受试者的未标准化 SMRW 权重记为 $w_i=\frac{e(X_i)}{1-e(X_i)}$。为了避免极端权重的影响，可以进一步进行标准化，标准化后的 SMRW 权重为 $w_i^*=\frac{w_i}{\sum w_i/N_h}$，其中 N_h 为外部对照组的受试者样本量。在“分析”阶段，Wang 等人采用了幂先验法，进而得到目标参数 θ 的幂先验分布表达式：

$$p(\theta|D_0,\delta)\propto\prod_{j=1}^{N_h}L(\theta|D_0)^{\delta\cdot w_j^*}\pi(\theta) \tag{13}$$

其中参数 δ 表示折扣系数。值得注意的是，当 $\delta=1$ 时 SMRW 权重实际上就是幂先验的折扣系数 $\delta\cdot w_j^*=w_j^*$，此时折扣系数在观察到结局事件前已经确定，因此不具有动态借用的性质。

② 数据适应性加权法

与 Wang 等人的方法相似，Harton 等人[27]同样在外部增强 RCT 设计下，提出了一种基于 On-trial 评分的“数据适应性加权(data adaptive weighting)”法。

在“设计”阶段，该方法首先从外部对照组受试者中挑选出 On-trial 评分最高的部分受试者，并与内部对照组受试者组成混合对照臂，进而模拟 1∶1 设计的 RCT 研究。对于混合对照臂中来自外部对照组的受试者，其权重为 $\hat{y}_i=\frac{e(X_i)}{1-e(X_i)}$，标准化后为 $\hat{y}_i^*=\frac{\hat{y}_i(N_{c_1}-N_{c_0})}{\sum_i^{N_{c_1}-N_{c_0}}\hat{y}_i}$，其中 N_{c_1}、N_{c_0} 分别为当前研究试验组和对照组的样本量；内部对照组中受试者的权重依旧为 1。在“分析”阶段，目标参数 θ 的幂先验分布表达式为：

$$p(\theta|D_0)\propto\prod_{j=1}^{N_{c_1}-N_{c_0}}L(\theta|D_0)^{\hat{y}_i^*}\pi(\theta) \tag{14}$$

同样，由于在观察到结局事件前折扣系数已经确定，因此也不具有动态借用的性质。

3. 相称性先验(dynamic borrowing)

除了幂先验法，Wang 等人[21]还将倾向性评分 SMRW 与相称性先验法相结合，并应用于外部增强 RCT 设计。其“设计”阶段与前述一致。以二分类结

局变量 Y 为例，在“分析”阶段先对目标参数进行 logit 变换，并假设转换后的参数服从正态分布，此时目标参数的相称性先验表达式同式(4)，其中似然函数为 $L(\theta_h \mid D_0) = \prod_{j \in N_h} \{\theta_h^{Y_j} (1 - \theta_h)^{1-Y_j}\}^{w_j^*}$，$N_h$ 为外部对照组的样本量，w_j^* 为标准化后的 SMRW 权重。

模拟研究

1. 不同偏倚模式下的统计学操作特性

一些研究通过模拟外部对照试验可能出现的偏倚，对“结合倾向性评分的贝叶斯外对照借用方法”的统计学操作特性进行评价。

(1)不存在任何偏倚

模拟研究显示，“结合倾向性评分的贝叶斯外对照组借用方法”能较好地控制 I 类错误膨胀与参数估计偏差；与不借用外部对照组相比，能够增加统计学效能，但与完全借用外部对照组相比，统计学效能有所损耗[21]。Wang 等人[21]发现单纯使用倾向性评分法也能够控制 I 类错误的膨胀、缩短置信区间的宽度并且提升统计效能。相比之下，“结合倾向性评分的贝叶斯外对照组借用方法”在可信区间宽度以及统计效能的提升上不及单纯使用倾向性评分法[21]。同样 Baron 等人[29]的模拟研究也显示，当基线协变量均衡可比且不存在未测量的混杂时，直接使用贝叶斯外对照借用方法的效果要优于“基于倾向性评分的贝叶斯逐个击破”分析框架。

(2)存在选择偏倚

模拟研究显示，此时“结合倾向性评分的贝叶斯外对照组借用方法”依旧能够控制参数估计的偏差和均方误差，以及 I 类错误的膨胀程度[21, 27-29]。与直接使用幂先验相比，Wang 等人[28]发现“倾向性评分分层结合幂先验法”的参数估计偏差和均方误差更小。Baron 等人[29]发现，使用“倾向性评分结合层次先验法”的参数估计偏差小于单纯使用层次先验。Harton 等人[27]发现，此时若使用固定系数幂先验法以及改良幂先验法，I 类错误将显著膨胀、参数估计的偏差也明显增加。相比之下，“数据适应性加权法”和“倾向性评分匹配结合幂先验法”的 I 类错误膨胀程度较小，参数估计偏差均变化不大[27]。同时在可信区间覆盖率上，后两者的表现也优于单独使用其他幂先验方法[27]。

(3)未测量的混杂因素

Wang 等人[21]发现，相比于单纯使用倾向性评分法，使用“倾向性评分结合贝叶斯的外对照组借用方法”在控制参数估计偏差以及 I 类错误上的表现更好，但是却损失了统计学功效。此外，模拟研究还显示，此时参数估计偏差增大的程度与协变量的个数并无关系[21]。

2. 倾向性评分与贝叶斯方法不同组合间的比较

一些研究还在不同偏倚模式下比较了倾向性评分法与贝叶斯外对照借用方法不同组合间的统计学操作性能。Wang 等人[21]发现，“倾向性评分加权结合幂先验法”对未测量混杂最为敏感，其次是“倾向性评分匹配和分层结合幂先验法”，而相称性先验相关的方法对未测量的混杂最不敏感[21]。随着混杂和偏倚的影响增大，“倾向性评分结合相称性先验法”的统计学效能和可信区间的覆盖概率高于其他方法，且参数估计偏差和 I 类错误的膨胀有限[21]。因此，Wang 等人推荐使用“倾向性评分加权或匹配结合相称性先验法”[21]。此外，Harton 等人[27]对比了“数据适应性加权法”与 Lin 等人的“倾向性评分匹配结合幂先验法”的统计学操作性能，结果显示二者的统计学性能相近。

讨　论

对于外部对照试验而言，保证当前研究与外部对照组受试者基线协变量的均衡可比十分重要。“结合倾向性评分的贝叶斯外对照借用方法”首先通过倾向性评分匹配、分层和加权等方法增加了研究间基线协变量的可比性，之后再使用贝叶斯外对照借用方法进行分析。模拟研究显示，当选择偏倚存在时，“结合倾向性评分的贝叶斯外对照借用方法”虽然在统计学功效方面有所损失，但是能够较好地控制了 I 类错误的膨胀与参数估计的偏差。

除了常见的倾向性评分匹配、分层以及加权外，Sachdeva 等人[33]在单臂(仅试验组)研究借用外部 RWD 数据进行单臂试验组扩增设计下，提出了一种基于倾向性评分的重要性抽样方法处理基线协变量的分布差异。虽然其在“分析”阶段使用的是频率学派的方法，但该方法也具备与幂先验折扣系数相结合的可能[33]。

目前“结合倾向性评分的贝叶斯外对照借用方法”存在几点不足。

首先，现有的方法在“设计”阶段就确定了外部对照组的信息借用程度，从而不具备动态借用的性质，如 Lin 等人提出的“倾向性评分结合幂先验法”。具有动态借用性质的贝叶斯外对照方法能够根据“先验——数据矛盾”，以数据驱动的方式客观地确定信息借用量，使得其不仅能较好地控制 I 类错误[13, 34]，且对未测量的偏倚不敏感，如结合倾向性评分的相称性先验[21]。因此，如何改造现有方法，使其具备动态借用的性质，是下一步研究需要考虑的问题。

其次，倾向性评分并不能保证组间基线协变量分布的完全均衡可比，特别是当倾向性评分模型指定错误或者存在未经测量的混杂因素时[26]。因此在使用

倾向性评分法后,需要对基线协变量分布的一致性进行再次评价[23-24]。然而现有方法对此重视不足,只有少数研究如 Wang 等人提出的“倾向性评分分层结合幂先验法”,通过计算各层内外部受试者间倾向性评分分布的重合度,实现对分层后基线协变量分布一致性的再评价。

考虑到完全正确指定倾向性评分模型存在困难,以及不能排除未测量混杂偏倚的影响,还可以采用其他处理基线协变量不可比问题的方法。Yu 等人[35]提出了基于熵权重的幂先验分析框架,该框架在“设计”阶段采用基于熵的协变量平衡法;在“分析”阶段,其将熵权重直接作为幂先验的折扣系数使用。此外,Jiang 等人[36]则是在弹性先验分析框架下,将基线协变量的分布差异融入其一致性工具的构建,以实现在基线协变量可比的条件下从外部对照组借用更多的信息量。

最后,除了选择偏倚和未测量的混杂偏倚外,目前模拟研究对于其他外可能发生的偏倚(如测量偏倚、错分偏倚、效应时间趋势等)关注较少。因此,有必要开展更为全面的模拟研究,对现有方法在不同偏倚模式下的统计学操作特性进行综合评价。

ICH E10 以及监管机构相关指导原则均强调了外部对照试验中偏倚控制的重要性[8-9, 37]。对于选择偏倚而言,本文介绍的“结合倾向性评分的贝叶斯外对照借用方法”可以较好地进行处理。但除了选择偏倚外,外部对照试验还可能发生其他偏倚,包括未测量混杂、效应时间趋势、测量和错分偏倚等[38]。然而,目前针对外部对照试验中其他偏倚的控制,方法学上仍存在空白。因此,如何从外部对照试验偏倚控制的角度出发,进一步构建和完善贝叶斯外对照借用方法学,将是未来研究的重点方向之一。

参 考 文 献

[1] FDA. Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products..(1998.05)[2023-04-01]. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products.

[2] 国家药品监督管理局药品审评中心. 成人用药数据外推至儿科人群的定量方法学指导原则(征求意见稿).(2022-09-19)[2023-04-01]. https://www.cde.org.cn/main/news/viewInfoCommon/765 8920dd1eb4cc37502d4880558aa9d.

[3] 国家药品监督管理局药品审评中心. 单臂临床试验用于支持抗肿瘤药上市申请的适用性技术指导原则.(2023-03-14)[2023-04-01]. https://www.cde.org.cn/main/news/viewInfoCommon/9f0c25dee6ba6781af809b36cf682eb6.

[4] 国家药品监督管理局药品审评中心. 罕见疾病药物开发中疾病自然史研究指导原则.(2022-01-06)[2023-04-01]. https://www.cde.org.cn/main/news/viewInfoCommon/c4e1ef312a0a0c039a7a4ca 55b91d4e8.

[5] 国家药品监督管理局药品审评中心. 药物真实世界研究设计与方案框架指导原则(试行).(2023-02-16)[2023-04-01]. https://www.cde.org.cn/main/news/viewInfoCommon/14aac16a4fc5b5841bc2529988a611cc.

[6] 国家药品监督管理局药品审评中心. 罕见疾病药物临床研发技术指导原则.(2022-01-06)[2023-04-01]. https://www.cde.org.cn/main/news/viewInfoCommon/c4e1ef312a0a0c039a7a4ca55b91d4e8.

[7] Pencina MJ, Thompson BT. Clinical Trials in the 21st Century — Promising Avenues for Better Studies. NEJM Evidence, 2022, 1(6): EVIDctw2200060.

[8] ICH. E10 Choice of Control Group and Related Issues in Clinical Trials.(2001-05)[2023-04-01]. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e10-choice-control-group-and-related-issues-clinical-trials.

[9] ICH. 临床试验中对照组的选择和相关问题 E10.(2000-07-20)[2023-04-01]. https://www.cde.org.cn/ichWeb/guideIch/toGuideIch/3/0.

[10] Jahanshahi M, Gregg K, Davis G, et al. The Use of External Controls in FDA Regulatory Decision Making. Therapeutic innovation & regulatory science, 2021, 55(5): 1019-1035.

[11] Lu N, Wang C, Chen WC, et al. Propensity score-integrated power prior approach for augmenting the control arm of a randomized controlled trial by incorporating multiple external data sources. J Biopharm Stat, 2022, 32(1): 158-169.

[12] Lin J, Gamalo-siebers M, Tiwari R. Ensuring exchangeability in data-based priors for a Bayesian analysis of clinical trials. Pharm Stat, 2022, 21(2): 327-344.

[13] Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharm Stat, 2014, 13(1): 41-54.

[14] Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. Environmetrics, 2006, 17(1): 95-106.

[15] Hobbs BP, Sargent DJ, Carlin BP. Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. Bayesian analysis, 2012, 7(3): 639-674.

[16] Pan H, Yuan Y, Xia J. A Calibrated Power Prior Approach to Borrow Information from Historical Data with Application to Biosimilar Clinical Trials. Journal of the Royal Statistical Society Series C, Applied statistics, 2017, 66(5): 979-996.

[17] Schmidli H, Gsteiger S, Roychoudhury S, et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics, 2014, 70(4): 1023-1032.

[18] Neuenschwander B, Capkun-niggli G, Branson M, et al. Summarizing historical information on controls in clinical trials. Clinical trials (London, England), 2010, 7(1): 5-18.

[19] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika, 1983, 70(1): 41-55.

[20] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate behavioral research, 2011, 46(3): 399-424.

(下转第 480 页)