arXiv:2602.07468v1 [stat.AP] 7 Feb 2026

存在协变量偏移时多区域临床试验区域治疗效应的一致性评估

Kunhai Qing[1],Xinru Ren[1], Jin Xu[1, 2, *]和 Menggang Yu[3, *]

[1] 华东师范大学统计学院，中国上海 200062[2] 统计与数据科学前沿理论及应用重点实验室 - 教育部，华东师范大学，中国上海 200062[3] 密歇根大学公共卫生学院生物统计系，美国密歇根州 48109

[*] 通信地址：jxu@stat.ecnu.edu.cn；menggang@umich.edu

摘要

多区域临床试验（MRCT）通过在全球不同人群中同步评估药物疗效和安全性，在新疗法开发中发挥着核心作用。评估各区域治疗效应的一致性是多区域临床试验的一项基本内容。现有方法通常关注区域特异性边际治疗效应。然而，当治疗效应异质性由效应修饰基线协变量引起时，这些协变量的分布差异可能导致错误结论。本文通过考虑条件平均治疗效应，在一致性评估中明确纳入这一现象。我们提出了一种两步评估策略，作为现有方法的补充，并减轻治疗效应异质性的影响。数值研究结果表明，所提出的方法有效。

关键词：条件平均治疗效应、协变量偏移、密度比、多区域临床试验、随机化

1 引言

随着药物研发日益全球化，多区域临床试验（MRCT）得到越来越广泛的应用，以加快患者入组，使患者更快获得治疗，促进协作与知识共享，并推动监管协调。多区域临床试验发展的一个重要里程碑，是国际人用药品注册技术协调会（ICH）于2017年11月发布了E17指南，题为《多区域临床试验计划与设计的一般原则》。随后，E17指南于2018年6月在日本获采纳并实施为监管要求，并于2018年7月在欧盟和美国获采纳并实施（Asano 等，2021）。

尽管多区域临床试验的总体结果被广泛视为监管决策的主要依据，但确保区域一致性仍至关重要。美国通常是制药公司的主要市场，因此美国参与者往往占多区域临床试验参与者的最大比例，通常超过 60%。因此，分配给其他区域的比例相对较小。日本药品医疗器械综合机构和中国国家药品监督管理局等监管机构，通常要求提供区域特异性证据，才能批准面向各自市场的药品。事实上，E17 强调样本量和统计分析计划的制定，以促进一致性评价。

然而，随着全球药物研发格局不断变化，对一致性进行评估的要求也已成为美国监管机构面临的一项重要议题。在美国食品和药物管理局于2024年9月16日发布的最新肿瘤学多区域临床试验（MRCT）开展FDA草案指南中（美国食品和药物管理局，2024年），美国食品药品监督管理局关注到“肿瘤学多区域临床试验中纳入的美国参与者比例递减”，这“可能限制对美国入组参与者的治疗效应与多区域临床试验总体研究人群中观察到的效应之间的一致性评估”。具体而言，指南指出了这样一种现象：“这些试验入组参与者的人口统计学特征或临床特征分布可能与

美国人群存在显著差异，因此外国数据可能不适合支撑美国食品药品监督管理局的监管决策”。

这种现象在机器学习（协变量偏移（Sugiyama 等人，2007）和因果推广（Chen 等人，2023a，b）中被称为这一术语。一个广为人知、并引起人们对这一问题高度关注的多区域临床试验（MRCT）实例是 PLATO（血小板抑制与患者结局）试验（Wallentin 等人，2009年）。这项 III 期研究表明，对于急性冠状动脉综合征（ACS）患者，替格瑞洛在预防心血管事件方面优于氯吡格雷。然而，根据预先设定的地区与治疗交互作用分析，北美地区的风险比却显示出相反的方向（Mahaffey 等人，2011）。仔细的事后分析发现，一个可能的原因是北美地区的试验受试者使用了较高剂量的阿司匹林。这两组分析（即预先设定分析与事后分析）的性质差异，导致美国食品药品监督管理局延迟批准替格瑞洛，这一结果是情有可原的（Armstrong，2021）。另一项多区域临床试验研究将用于我们的数值分析，即 BELIEVE 试验。该试验支持美国食品药品监督管理局于 2019年11月批准罗特西普，用于治疗需要定期接受红细胞输注的成年患者的贫血，患者患有β-地中海贫血（Cappellini 等人，2018年）。基线输血需求较低的患者获得了更强的治疗效应（Sheth 等人，2023）。因此，在中国，罗特西普仅获批用于需要定期接受红细胞输注，且每 24 周需要 ≤ 15 单位的患者。

E17指南指出，描述性汇总、图形展示、协变量调整分析以及治疗与区域交互作用检验，均可作为此类评估的可能策略。尽管这些策略在探索和解释方面非常有用，但它们并未提供用于一致性判定的预先规定规则，尤其是在存在协变量偏移的情况下。

几乎所有关于一致性评价预先指定程序的现有文献，都建立在日本厚生劳动省（MHLW）于2007年9月发布的白皮书《全球临床试验基本原则》所倡导的两种方法之上（日本厚生劳动省，2007）。具体而言，第一种方法假设总体效应具有统计学显著性，并通过计算特定区域内的治疗效应超过总体效应某一比例的似然来评价一致性，而总体效应在统计学上具有

显著性。第二种方法计算所有区域的治疗效应方向一致的概率。两种方法都提供了用于判定一致性的预先指定阈值，因此可用于区域分配的样本量确定，也可作为计划分析。它们确实广泛应用于多区域临床试验中（Asano 等，2021）。

MHLW 指南发布后，研究者陆续开发了多种方法（Ko 等，2010；Ikeda 和 Bretz，2010；Tsou 等，2011；Chen 等，2012；Tsong 等，2012；Wu 等，2020），用于评估区域一致性并确定区域样本量要求。这些方法假设各区域具有一致的潜在治疗效应，且方差相同。为更灵活地对区域差异进行建模，研究者开发了一种纳入区域特异性治疗效应的扩展框架，使 MHLW 指南（日本厚生劳动省，2007）所述标准得以修改和扩展，并推动了其他一致性评价方法的提出（Chen 等，2010；Quan 等，2010）。此外，研究者还提出了基于假设检验的方法（Tanaka 等，2012；Teng 等，2017；Li 等，2024）和随机效应模型（Quan 等，2013）来评估一致性。

目前，针对多区域临床试验中的一致性评价，尚无预先指定程序能够明确处理治疗效应修饰变量可能出现的分布偏移，本文也将其称为协变量偏移。在第 2 节中，我们首先重新审视日本厚生劳动省型评估方法（日本厚生劳动省，2007），这类方法直接使用感兴趣区域的平均治疗效应估计值进行一致性评估。我们将其称为单步方法，以区别于第 3 节提出的方法；后者引入了额外步骤，以便在治疗效应修饰变量发生分布偏移时进行一致性评估。第 4 节通过模拟研究评估所提方法的性能，并展示其操作特征。第 5 节通过一个真实案例说明所提方法的实际效用。第 6 节总结研究发现，并讨论未来研究方向。

2 一步一致性评估

我们首先介绍符号，并描述日本厚生劳动省型评估方法（日本厚生劳动省，2007）；该方法直接使用感兴趣区域的平均处理效应估计值进行一致性评估。我们提出的方法纳入了一个额外步骤，以便在治疗效应修饰变量发生分布偏移时进一步评估一致性。

令 $T \in \{-1, 1\}$ 表示治疗指示变量，其中 $T = 1$ 对应治疗组，$T = -1$ 对应对照组。我们假设 T 经过随机分配，且 $\Pr(T = 1) = \pi_1$，$\Pr(T = -1) = \pi_{-1} = 1 - \pi_1$。令 Y 表示感兴趣的终点，令 $\boldsymbol{X} = (X_1, \cdots, X_p)^\top$ 表示一个 p 维的基线协变量向量，该向量在治疗分配前测量。条件平均治疗效应（CATE）定义为 $\Delta(\boldsymbol{X}) = \mathbb{E}(Y \mid T = 1, \boldsymbol{X}) - \mathbb{E}(Y \mid T = -1, \boldsymbol{X})$，而 ATE 是无条件期望，表示为 $\delta = \mathbb{E}\{\Delta(\boldsymbol{X})\} = \mathbb{E}(Y \mid T = 1) - \mathbb{E}(Y \mid T = -1)$。

假设该多区域临床试验涉及 K 个区域。令 R 表示区域。对于每个区域$r = 1, \cdots, K$，令 $F_r(\cdot) = P(\boldsymbol{X} \le \cdot \mid R = r)$ 和 $F_{-r}(\cdot) = P(\boldsymbol{X} \le \cdot \mid R = -r)$ 分别表示协变量向量 $\boldsymbol{X}$ 在区域 r 内以及区域 r 之外（或互补区域）的分布函数，其中此处的不等式按分量解释。区域特异性条件平均治疗效应为

$$\Delta_r(\boldsymbol{X}) = \mathbb{E}(Y|T = 1, \boldsymbol{X}, R = r) - \mathbb{E}(Y|T = -1, \boldsymbol{X}, R = r)$$

and

$$\Delta_{-r}(\boldsymbol{X}) = \mathbb{E}(Y|T = 1, \boldsymbol{X}, R = -r) - \mathbb{E}(Y|T = -1, \boldsymbol{X}, R = -r).$$

其对应的平均处理效应为

$$\delta_r = \mathbb{E}_r\{\Delta_r(\boldsymbol{X})\} \quad \text{and} \quad \delta_{-r} = \mathbb{E}_{-r}\{\Delta_{-r}(\boldsymbol{X})\},$$

其中，期望是相对于区域分布 F_r 和 F_{-r}计算的。

令 $\{(y_i, \boldsymbol{x}_i, t_i, r_i) : i = 1, \cdots, n\}$ 为从 $(Y, \boldsymbol{X}, T, R)$ 中抽取的容量为 n 的简单随机样本，而 $\mathcal{D}_r$ 和 $\mathcal{D}_{-r}$ 分别表示区域样本 $\{(y_i, \boldsymbol{x}_i, t_i, r_i = r)\}$ 以及互补

区域样本 $\{(y_i,\ \boldsymbol{x}_i,\ t_i,\ r_i \neq r)\}$。对于 $r = 1,\ \cdots,\ K$，令 n_r 表示区域 r 的样本量。因此，总样本量为 $n = n_1 + \cdots + n_K$。令 $\rho_r = n_r/n$ 表示区域比例。为简化符号，我们假设

$$\frac{\sum_{i=1}^{n} I(t_i = 1)I(r_i = r)}{\sum_{i=1}^{n} I(t_i = 1)} = \frac{\sum_{i=1}^{n} I(t_i = -1)I(r_i = r)}{\sum_{i=1}^{n} I(t_i = -1)} = \rho_r.$$

这一条件对于我们的方法学开发并非必要。总体平均治疗效应的经验估计量为

$$\widehat{\delta} = \frac{\sum_{i=1}^{n} y_i I(t_i = 1)}{\sum_{i=1}^{n} I(t_i = 1)} - \frac{\sum_{i=1}^{n} y_i I(t_i = -1)}{\sum_{i=1}^{n} I(t_i = -1)}.$$

$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$ 令 $Z = \delta/\widehat{\ }$se(δ)，其中 $\widehat{\ }$se(δ) 表示 δ的标准误。我们考虑一项优效性试验，其中响应值越高越理想。当 $Z > z_\alpha$时，研究者应拒绝关于总体治疗效应的原假设$\widehat{\ }$（H_0：$\delta = 0$），其中 α 为名义显著性水平，z_α 为标准正态分布的上 α 百分位数。

区域 r 以及区域外 r 的平均处理效应经验估计量分别为

$$\widehat{\delta}_r = \frac{\sum_{i=1}^{n} y_i I(t_i = 1)I(r_i = r)}{\sum_{i=1}^{n} I(t_i = 1)I(r_i = r)} - \frac{\sum_{i=1}^{n} y_i I(t_i = -1)I(r_i = r)}{\sum_{i=1}^{n} I(t_i = -1)I(r_i = r)},$$
$$\widehat{\delta}_{-r} = \frac{\sum_{i=1}^{n} y_i I(t_i = 1)I(r_i \neq r)}{\sum_{i=1}^{n} I(t_i = 1)I(r_i \neq r)} - \frac{\sum_{i=1}^{n} y_i I(t_i = -1)I(r_i \neq r)}{\sum_{i=1}^{n} I(t_i = -1)I(r_i \neq r)}.$$

传统一致性标准认为，当区域平均处理效应 δ_r 等于（固定效应模型）或近似等于（随机效应模型，在共同先验分布下）总体平均处理效应 δ，或区域 r 以外的平均处理效应 δ_{-r} 时，区域治疗效应与总体治疗效应一致。与正式的假设检验框架不同，日本厚生劳动省型评估方法（日本厚生劳动省，2007）侧重于确保事件 $\{\delta_r > q\delta\}$（Quan等人，$\widehat{\ }$ $\widehat{\ }$2010）或 $\{\delta_r > q\delta_{-r}\}$（Ko等，2010）发生的概率较高，其中 $q \geq 0.5$ 是需要保留的总体治疗效应比例。这也称为“选择设计”（Gibbons等人，1999；Chen和Yu，2024；Sargent和Goldberg，2001）。这些标准仅依赖边际治疗效应，不考虑协变量；因此，我们将其归类为边际一致性标准。

监管机构主要要求多区域临床试验证明总体成功。只有在达到总体统计学显著性后，才会考虑一致性。因此，将一致性评估为条件事件既合理又恰当。在实践中，许多一致性评估方法都遵循这一原则，仅在总体试验达到统计学显著性时评估一致性；参见Kawai等人(2008)、Ko等人(2010)、Wu等人(2020）及其中的参考文献。因此，我们可以将上述事件改写为

$$\left\{\frac{\widehat{\delta}_r}{\widehat{\delta}} > q \,\middle|\, \widehat{Z} > z_\alpha\right\} \text{ and } \left\{\frac{\widehat{\delta}_r}{\widehat{\delta}_{-r}} > q \,\middle|\, \widehat{Z} > z_\alpha\right\}, \tag{1}$$

后者等同于

$$\left\{\frac{\widehat{\delta}_r}{\widehat{\delta}} > \frac{q}{1-\rho_r+q\rho_r} \,\middle|\, \widehat{Z} > z_\alpha\right\}.$$

一旦发生这一条件事件，即认为满足一致性。下文中，我们将用第二个事件来解释我们的方法。

3 两步一致性评估

对于任意特定区域 r，区域平均处理效应定义为 $\delta_r = \mathbb{E}_r\{\Delta_r(\boldsymbol{X})\} = \int \Delta_r(\boldsymbol{x})dF_r(\boldsymbol{x})$。这一表达式指出，区域 r 与总体人群（或区域 r 之外的区域）之间的平均处理效应差异可能源于：(i) 条件平均处理效应存在显著差异，即$\Delta_r(\boldsymbol{X}) \neq \Delta_{-r}(\boldsymbol{X})$；(ii) 协变量分布存在显著差异，即 $F_r(\boldsymbol{X}) \neq F_{-r}(\boldsymbol{X})$；或 (iii) 两种因素共同作用。

因此，当一步评估提示可能存在不一致性时，即当$\frown \frown \delta_r/\delta \leq q$，进一步调查其根本原因就十分重要。具体而言，应评估这种不一致性是源于条件平均处理效应的差异、协变量分布的差异，还是两者兼有。为此，我们提出以下通用算法：

（i）进行单步一致性评估。如果结论为一致，则停止；

(ii) 否则，如果结论是不一致性，则评估 $\Delta_r(\boldsymbol{X})$ 和 $\Delta_{-r}(\boldsymbol{X})$ 是否彼此接近；

(ii.1) 如果 $\Delta_r(\boldsymbol{X})$ 和 $\Delta_{-r}(\boldsymbol{X})$ 彼此不接近，则最终结论为不一致性；停止；

(ii.2) 如果 $\Delta_r(\boldsymbol{X})$ 与 $\Delta_{-r}(\boldsymbol{X})$ 彼此接近，则估计区域 r 的协变量偏移调整后的平均处理效应，并重新评估一致性；

(ii.3) 随后，将基于 (ii.2) 得出的结论作为最终结论。

该算法通过基于步骤 (ii) 的补救策略扩展了一步法。步骤 (ii) 的依据——评估 $\Delta_r(\boldsymbol{X})$ 与 $\Delta_{-r}(\boldsymbol{X})$ 是否相似 —— 如下。条件平均处理效应和协变量分布的差异可能相互抵消。也就是说，即使 $\Delta_r(\boldsymbol{X}) \neq \Delta_{-r}(\boldsymbol{X})$ 和 $F_r(\boldsymbol{X}) \neq F_{-r}(\boldsymbol{X})$ 均成立，$\delta_r = \delta_{-r}$ 仍有可能成立。因此，在估计区域 r 的协变量偏移调整后的平均处理效应之前，有必要评估 $\Delta_r(\boldsymbol{X})$ 与$\Delta_{-r}(\boldsymbol{X})$ 是否相似。否则，不宜采用补救策略。具体步骤如下。

3.1 评估 Δ_r（$\boldsymbol{X}$）与 Δ_{-r}（$\boldsymbol{X}$）之间的相似性

我们通过估计个体处理效应（ITE），来评估 Δ_r（$\boldsymbol{X}$）与 Δ_{-r}（$\boldsymbol{X}$）是否彼此接近。具体而言，很容易看出（LiangandYu，2022；Chen 等人，2017年）

$$\mathbb{E}\left[\frac{T_i}{\pi_{T_i}} Y_i \,\middle|\, \boldsymbol{X}_i, R_i = r\right] = \Delta_r(\boldsymbol{X}_i),$$

其中 $\pi_{T_i} = \pi_1$ 当 $T_i = 1$ 且 $\pi_{T_i} = \pi_{-1}$ 当 $T_i = -1$。此外，对于任意函数m（$\boldsymbol{X}_i$，R_i），

$$\mathbb{E}\left[\frac{T_i}{\pi_{T_i}}\{Y_i - m(\boldsymbol{X}_i, R_i)\} \,\middle|\, \boldsymbol{X}_i, R_i = r\right] = \Delta_r(\boldsymbol{X}_i) \tag{2}$$

因此，${\pi_{T_i}}^{-1} T_i\{Y_i - m(\boldsymbol{X}_i, R_i)\}$ 可视为个体处理效应的估计值。因此，我们考虑以下工作模型

$$\frac{T_i}{\pi_{T_i}}\{Y_i - m(\boldsymbol{X}_i, R_i)\} = \beta_0 + \boldsymbol{\beta}_X^\top \boldsymbol{X}_i + \beta_R \cdot I(R_i = r) + \boldsymbol{\beta}_{RX}^\top \boldsymbol{X}_i \cdot I(R_i = r) + \varepsilon_i. \tag{3}$$

Wu 和 Gagnon-Bartsch (2018年指出，对于 $m(\boldsymbol{X}_i,\ R_i)$，最有效的选择是 $\pi_{-1}\mathbb{E}[Y_i \mid T_i = 1, \boldsymbol{X}_i, R_i] + \pi_1\mathbb{E}[Y_i \mid T_i = -1, \boldsymbol{X}_i, R_i]$。由于期望值未知，因此需要对其进行估计。Wu 和 Gagnon-Bartsch(2018年) 提出了一种留一法潜在结果（LOOP）估计量。该方法利用其他参与者的协变量信息（$\boldsymbol{x}_j,\ j \neq i$）来提供估计值 $\widehat{m}_i$，从而显著降低方差。此外，LOOP 估计量具有稳健性，因为它不依赖任何模型假设；考虑到来自区域 r 的样本量通常较小，这一特点尤为有益。由于有偏的 $\widehat{m}_i$ 可用于（3），且 $\widehat{m}_i$ 的收敛速率不必达到 $\sqrt{n}$，从业者可以使用线性回归等简单模型，或随机森林等复杂的机器学习方法（Breiman，2001）进行估计。

–

最后，由于 $\boldsymbol{\beta}_{RX}$ 是一个向量，我们使用Wald检验或似然比检验来检验其显著性。拒绝 $H_0:\ \boldsymbol{\beta}_{RX} = \mathbf{0}$表明 $\Delta_r(\boldsymbol{X}) \neq \Delta_{-r}(\boldsymbol{X})$，也就是说，区域 r 与区域外 r之间，协变量 $\boldsymbol{X}$ 与终点 Y交互作用存在显著差异。否则，我们认为这两个条件平均处理效应函数相似。

3.2 协变量分布偏移调整后的平均处理效应

当 $\Delta_r(\boldsymbol{X})$和 $\Delta_{-r}(\boldsymbol{X})$非常相似时，平均处理效应 δ_r 与 δ_{-r}之间的较大差异主要应源于处理效应修饰因子的分布差异。基于条件平均处理效应的一致性应能导出平均处理效应一致性结论这一假设，我们考虑以下量：

$$\delta_r^* \ = \ \mathbb{E}_r\left\{\Delta_r(\boldsymbol{X}) \cdot \frac{f_{-r}(\boldsymbol{X})}{f_r(\boldsymbol{X})}\right\} = \mathbb{E}_{-r}\left\{\Delta_r(\boldsymbol{X})\right\}.$$

因此，δ_r^* 是根据区域 $-r$的协变量分布边际化后的区域 r 条件平均处理效应。下文将论证，δ_r^* 可与 δ_{-r}直接比较，后者是

互补区域的平均处理效应，用于评估治疗效应的一致性，

$$\begin{aligned}\delta_r^* &= \mathbb{E}_r\left\{\Delta_r(\boldsymbol{X})\cdot\frac{f_{-r}(\boldsymbol{X})}{f_r(\boldsymbol{X})}\right\}\\ &= \int \Delta_r(\boldsymbol{x})\cdot\frac{f_{-r}(\boldsymbol{x})}{f_r(\boldsymbol{x})}f_r(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \int \Delta_{-r}(\boldsymbol{x})\cdot f_{-r}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= \mathbb{E}_{-r}\left\{\Delta_{-r}(\boldsymbol{X})\right\} = \delta_{-r}\end{aligned}$$

δ_r^* 的估计可以基于

$$\widehat{\delta}_r^* \quad = \quad n_r^{-1}\sum_{i\in\mathcal{D}_r}\widehat{\Delta}_r(\boldsymbol{x}_i)\cdot\frac{\widehat{f}_{-r}(\boldsymbol{x}_i)}{\widehat{f}_r(\boldsymbol{x}_i)} \tag{4}$$

^其中 $\Delta_r(\boldsymbol{x}_i)$ 是个体处理效应的估计值。具体而言，可以使用^ ^第 3.1 节中介绍的 LOOP估计量。密度比估计 $f_{-r}(\boldsymbol{x}_i)/f_r(\boldsymbol{x}_i)$ 可以使用核平滑方法获得（Wand和Jones，1994年）。

^ ^当维度 p 中等偏大时，密度比估计 $f_{-r}(\boldsymbol{x}_i)/f_r(\boldsymbol{x}_i)$ 可能不稳定。解决这一问题的一种策略是采用充分降维（Liang和Yu，2022）。具体而言，当 $\Delta_r(\boldsymbol{x}_i)$ 仅取决于少数几个线性组合 $\boldsymbol{x}_i^d \equiv \boldsymbol{\alpha}^\top\boldsymbol{x}_i$ 时，其中 $\boldsymbol{\alpha}$ 是维度为 $p\times d$ 的矩阵，且 d 较小时，可以在式（4）中使用两个区域之间 $\boldsymbol{x}_i^d$ 的密度比。甚至可以固定 $d=1$，并使用单指标模型（Xiong等人，2017年）估计 $\Delta_r(\boldsymbol{x}_i)$。

然而，线性组合可能不易解释。因此，我们考虑以下替代方案。对于 $\boldsymbol{x}$ 的每个分量 x_s $s=1,\ \cdots,\ p$，我们将新估计量定义为：

$$\widehat{\delta}_{r,s}^* = n_r^{-1}\sum_{i\in\mathcal{D}_r}\widehat{\Delta}_r(\boldsymbol{x}_i)\cdot\frac{\widehat{f}_{-r,s}(x_{s,i})}{\widehat{f}_{r,s}(x_{s,i})}$$

^我们可以看到，$\delta^*_{r,\ s}$ 收敛至

$$\begin{aligned}
&\mathbb{E}_r\left\{\Delta_r(\boldsymbol{X})\cdot\frac{f_{-r,s}(X_s)}{f_{r,s}(X_s)}\right\}\\
&=\int\Delta_r(\boldsymbol{x})\cdot\frac{f_{-r,s}(x_s)}{f_{r,s}(x_s)}\cdot f_r(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&=\int\Delta_r(\boldsymbol{x})\cdot f_r(\boldsymbol{x}_{-s}\mid x_s)\cdot f_{-r,s}(x_s)\mathrm{d}\boldsymbol{x}\\
&=\int\left\{\int\Delta_r(\boldsymbol{x})\cdot f_r(\boldsymbol{x}_{-s}\mid x_s)\mathrm{d}\boldsymbol{x}_{-s}\right\}\cdot f_{-r,s}(x_s)\mathrm{d}x_s
\end{aligned}\tag{5}$$

其中 $\boldsymbol{x}_{-s}$ 为 是 $\boldsymbol{x}$ 中除去 x_s后的组成部分。上述大括号中的量为

$$\mathbb{E}_r\big\{\Delta_r(\boldsymbol{X})\mid X_s=x_s\big\}=\Delta_r(x_s),$$

其中

$$\Delta_r(x_s)\equiv\mathbb{E}(Y|T=1,X_s=x_s,R=r)-\mathbb{E}(Y|T=-1,X_s=x_s,R=r)$$

是仅以 X_s为条件的治疗效应。

因此，(5) 等于

$$\int\Delta_r(x_s)f_{-r,s}(x_s)\mathrm{d}x_s=\mathbb{E}_{-r}\big\{\Delta_r(X_s)\big\}\equiv\delta^*_{r,s}.$$

在各区域的条件平均处理效应相似这一假设下，如果进一步假设 $f_r(\boldsymbol{x}_{-s}|x_s)=f_{-r}(\boldsymbol{x}_{-s}|x_s)$，则 $\Delta_r(x_s)$和 $\Delta_{-r}(x_s)$也应当相似。因此， $\mathbb{E}_{-r}\{\Delta_r(X_s)\}$将与 $\mathbb{E}_{-r}\{\Delta_{-r}(X_s)\}=\delta_{-r}$ 相似。此性质使$\delta^*_{r,s}$ 成为我们进行评估的合适目标。在计算出$\delta^*_{r,s}$后，对于 $s=1,\cdots,p$, 我们提出对式 (1) 中第二个条件事件的修订版本，如下所示:

$$\left\{\max_s\left(\frac{\widehat{\delta}^*_{r,s}}{\widehat{\delta}_{-r}}\right)>q^{(2)}\,\middle|\,\widehat{Z}>z_\alpha\right\},\tag{6}$$

其中， $q^{(2)}$ 是第二步中的阈值。

令 $q^{(1)}$ 表示总体治疗效应中要保留的比例，以用于所提出两步法的第一步。按照设计，当 $q^{(1)}=q$ （即与一步法相同的阈值）时，无论 $q^{(2)}$取何值，我们提出的两步法所得覆盖概率都不低于一步法。借鉴假设检验的方法，我们寻求一种

算法1： 两步一致性评估

输入： α，$q^{(1)}$，$q^{(2)}$，$\mathcal{D}_r$和 $\mathcal{D}_{-r}$

输出： “一致性”或“不一致性”

计算 $\widehat{Z}, \widehat{\delta}_r, \widehat{\delta}_{-r}$ 和 $\widehat{\delta}^*_{r,s}\ s = 1, \cdots, p$;

if $\widehat{Z} > z_\alpha$ 和 $\widehat{\delta}_r / \widehat{\delta}_{-r} > q^{(1)}$ （即边际一致性评估）**则**

  返回 “一致性”;

否则

  如果 H_0： $\Delta_r(\boldsymbol{X}) = \Delta_{-r}(\boldsymbol{X})$ 被拒绝 **则**

    返回 “不一致性”;

  else

    如果 至少有一个不等式 $\widehat{\delta}^*_{r,s} / \widehat{\delta}_{-r} > q^{(2)}, s = 1, \cdots, p,$ 成立 **则**

      返回 “一致性”;

      else

        返回 “不一致性”;

将 $q^{(1)}$ 与 $q^{(2)}$ 结合起来的方式，使得在不存在协变量偏移时，若条件平均处理效应函数不同，其覆盖概率低于一步法（类似于第一类错误）；若条件平均处理效应函数相似，其覆盖概率则高于一步法（类似于统计功效）。因此，我们有意选择一个大于 $q^{(1)}$ 的 q 值，以高置信度保证第一步所声称的一致性，并使用 $q^{(2)} = q$ 来纠正因实际发生协变量偏移而导致的错误不一致性判断。我们在第4节通过模拟探讨 $q^{(1)}$ 的不同取值。

3.3 两步评估算法及实践考量

用于确定区域之间一致性的两步评估完整流程 r以及该区域与其他区域之间的一致性，详见算法1。

算法中用于声称一致性的增广步骤可能引发这样的担忧：

当第 3 节开头所述的情景不成立时，该方法可能会增加假阳性结论，具体而言，即不存在协变量偏移，或 CATE 不依赖于某个特定协变量时 X_s。在此我们从理论角度论证，这种担忧并无依据。

首先，如果某个协变量 X_s不存在分布偏移，则密度比$f_{-r,\ s}/f_{r,\ s}$ 等于 1。因此，估计目标 $\delta^*_{r,\ s}$ 简化为 δ_r。同样，当 $\Delta_r(x_s)$ 不依赖于 x_s 时，它等于 δ_r ，这也意味着 $\delta^*_{r,\ s} = \delta_r$。在这两种情况下，$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$增广步骤中使用的比值 $\delta^*_{r,\ s}/\delta_{-r}$ 将非常接近第一步中的比值 δ_r/δ_{-r} 。因此，传统一步法的操作特征得以保留。我们的模拟研究也支持这一现象。

另一方面，如果对于某个 $s = 1, \cdots, p$，不等式 $\widehat{\delta}^*_{r,\ s}/\widehat{\delta}_{-r} > q^{(2)}$ 成立，我们可以得出结论：观察到的区域平均处理效应与互补区域平均处理效应之间的差异，是由该特定协变量的分布偏移造成的。这有助于识别起作用的协变量（即效应修饰变量），并提高研究结果的可解释性。因此，比值集合 $\{\delta^*_{r,s}/\delta_{-r}, s = 1, \cdots, p\}$ 能够提供有关各个协变量对平均处理效应差异影响的有用信息。对这些比值进行排序有助于分析，也可作为整体一致性评估的补充可视化工具，这与 E17指南的要求一致。

在多区域临床试验中，解释条件平均处理效应可能具有挑战性，尤其是在协变量为连续变量时。一种常见且有效的方法是将连续协变量离散化为分类变量。通过这种转换，未知函数 $\Delta(\boldsymbol{X})$ 会变为分段常数，从而显著提高后续统计分析（如协方差分析）的可解释性和有效性。因此，在应用我们提出的一致性评估方法时，我们建议对协变量进行离散化，这也能简化密度比的估计。

尽管我们提出的方法可直接适用于连续和二元结局，但在处理至事件时间终点时会出现重大的方法学挑战。条件平均处理效应框架不能直接应用于生存数据。此外，一种常见的替代方法——条件风险比——给一致性评估带来困难，因为其

期望通常并不等于总体风险比。

为规避这一局限性，我们采用限制性平均生存时间（限制平均生存时间）（Pepe 和 Fleming，1989年）作为该情境下治疗效应更合适的主要指标。限制平均生存时间衡量直至具有临床意义的截断点 τ的生存时间，并可作为风险比的有力替代指标（Uno 等，2014年）。为解决生存数据固有的右删失问题，我们将删失生存时间转换为伪观测值（Andersen 和 Perme，2010年），用于限制平均生存时间估计；详细步骤见附录（A1节）。该转换可使用R包 `pseudo`实现。通过将伪观测值作为结局，我们有效地将生存时间终点转化为连续响应。因此，我们可以应用所提出的方法进行一致性评估。

4 模拟

我们开展大量模拟，以评估所提出方法在不同终点类型下的表现，并与Ko等人（2010年）的一步法进行比较。

4.1 设置

将感兴趣区域的样本量设为 $n_r = 60$，互补区域的样本量设为 $n_{-r} = 340$，由此得到 $\rho_r = 0.15$。这一总体样本量和区域比例代表了实践中的常见情形。我们考虑四个协变量，即 $p = 4$。

为模拟真实试验场景，我们假设 $\boldsymbol{X}$ 的支撑集有界。具体而言，每个分量 X_s（$s = 1$，2，3，4）均独立地取自如下截断正态分布

$$f(x|a,b,\mu,\sigma) = \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\sigma\left\{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right\}}, \quad a \le x \le b,$$

其中，Φ 和 ϕ 分别是标准正态变量的分布函数和密度函数。在所有模拟中，截断界固定为 $a =$

-3 和 $b = 3$，标准差设为 $\sigma = 1.4$。为了引入感兴趣区域 r 的样本与其互补区域之间的分布偏移，我们为与区域 r 和$-r$相关的某些协变量设定不同的参数 μ 值。

潜在结果 $Y_h(t)$ 在区域 $h(h = r, -r)$ 中具有 $t \in \{-1, 1\}$ 的底层模型规定如下。对于区域 h 中的连续型结局，假设

$$Y_h(t) = 10 + 5\sum_{s=1}^{4} X_s + \kappa_h g(X_1, X_2)\frac{t+1}{2} + \varepsilon,$$

其中 $\varepsilon \sim N(0, 1)$，且 g 是仅取决于 X_1 和 X_2 的已知函数。因此，区域 h 的条件平均处理效应函数为 $\Delta_h(\boldsymbol{X}) = \kappa_h g(X_1, X_2)$，其中 κ_h 控制条件平均处理效应的尺度。对于 g，我们考虑三种形式：(i) 线性：$X_1 + 0.5X_2$；(ii) 二次：$X_1^2 - 1.61 + 0.5(X_2^2 - 1.61)$；(iii) 三次：$X_1^3 + 0.5X_2^3$。

对于二元结局，区域 h 中成功概率经逻辑斯蒂变换后的值为

$$\operatorname{logit} \mathbb{P}\{Y_h(t) = 1 \mid \boldsymbol{X}\} = -5 + 5(X_3 + X_4) + \kappa_h(X_1 + 0.5X_2)\frac{t+1}{2}.$$

对于生存结局，假设区域 h 中的事件发生时间服从指数分布，具体如下

$$Y_h(t) \sim \operatorname{Exp}(\lambda_{h,t}) \text{ with } \lambda_{h,t} = X_3 + X_4 + \kappa_h(X_1 + 0.5X_2)\frac{t+1}{20}.$$

删失时间服从均匀分布 $U[0, 200]$，，由此得到的删失率约为 25%。如第 3.3 节末尾所述，我们将使用 伪观测值 (Andersen 和 Perme，2010年) 来处理删失，并基于 $Y^* = \min(Y, \tau)$ 进行限制平均生存时间估计，其中限制时间 τ设为 100。请注意，对于二元和生存终点两种情形，区域 h的条件平均处理效应函数都依赖于所有协变量，其解析形式的详细信息见附录。参数 κ_h 与连续型终点情形一样，用于控制条件平均处理效应的尺度。

为引入区域间条件平均处理效应的差异，我们固定 $\kappa_{-r} = 10$，并在集合 $\{0, 2, 4, 6, 8, 10\}$上变化κ_r，从而使比值 κ_r/κ_{-r}跨越离散

范围 $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$。比值为零表示区域间条件平均处理效应的差异最大，比值为一表示区域间条件平均处理效应相同。

X_s（$s = 1$、2、3、4）均根据各自的经验 33% 和 66% 分位数，等距离散为三个水平。处理变量 T 服从成功概率为 0.5 的伯努利分布。

令 $\text{CP}^{\text{Ko}}(q) = \Pr(\widehat{\delta}_r/\widehat{\delta}_{-r} > q|\widehat{Z} > z_\alpha$表示使用传统一步法计算的区域 r相对于其互补区域的一致性概率（CP）（Ko等人，2010年）。令 $\text{CP}^{\text{ts}}(q^{(1)}, q^{(2)}) = \Pr$（算法1 声称一致$|\widehat{Z} > z_\alpha$）表示使用我们提出的两步法得到的相同判据的一致性概率。令 $q = q^{(2)} = 0.5$并考虑 $q^{(1)}$ 取 0.5、0.75 和 0.9，正如第3.2节末尾所提出的。整个模拟重复 $B = 10,000$ 次。

4.2 不存在协变量分布偏移时的结果

我们首先展示不存在协变量分布偏移时的模拟结果。也就是说，对于所有 $s = 1$、.....、4，各区域的协变量分布均相同，其意义为 $F_r(X_s) = F_{-r}(X_s)$（或 $\mathbb{E}_r(X_s) = \mathbb{E}_{-r}(X_s)$ 的较弱版本）。

为引入总体显著性的平均处理效应，我们固定 $\mu = 0$，令 X_2、X_3 和 X_4，并调整参数 μ，令 X_1，使得在所考虑的 κ_r 和 κ_{-r} 组合下，多区域临床试验（通过 t 检验）的统计功效均大于 90%，因为统计功效越高，就越容易证明我们提出的方法能够提高覆盖概率。附录中的表A1 列出了连续型终点中 X_1 分别用于 g 的线性、二次和三次形式时所得的矩及相应的区域平均处理效应；表 A2 则列出了二分类与生存终点中 X_1 所得的一阶矩。由于协变量分布相同，区域平均处理效应 δ_h除尺度参数 κ_h外均相同。

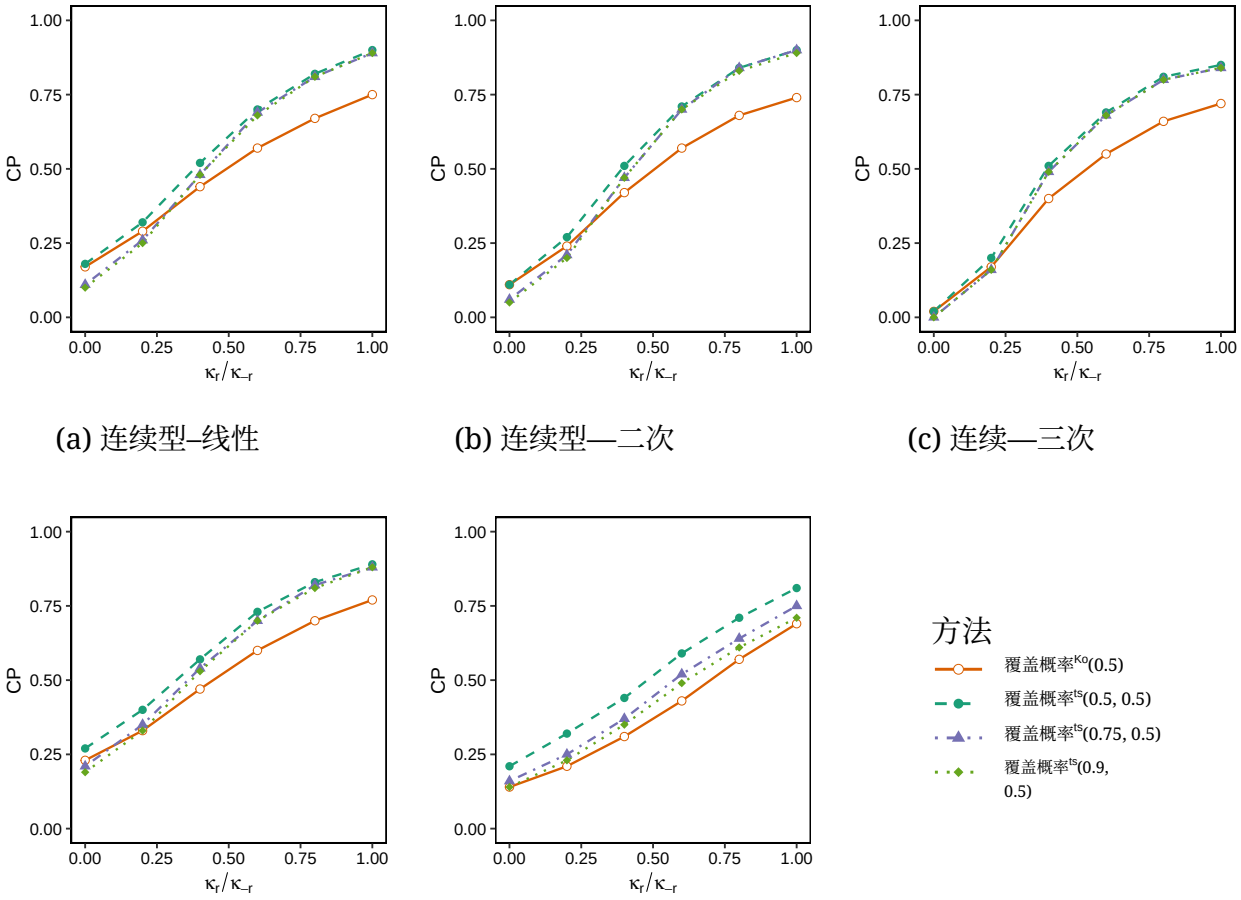
在这些设定下，图1展示了区域 r 相对于其互补区域的经验 CP。随着比值 κ_r/κ_{-r} 增大，区域CATE函数之间的差异逐渐减小。因此，两种方法得到的 CP 均

相应增大。这一趋势符合我们的预期，因为在不存在协变量分布偏移的情况下，CATE函数差异是不一致性的主要驱动因素。总体而言，在所有终点上，两步法的表现都略优于一步法。当CATE之间的差异较大时（比值 κ_r/κ_{-r} 较小，例如小于0.4），两步法得到的经验 CP 通常较低；当CATE之间的差异较小时（比值 κ_r/κ_{-r} 较大，例如大于0.6），两步法得到的经验 CP 通常较高。结果表明，与Ko等人(2010年)提出的传统一步法相比，在存在协变量偏移时，我们提出的两步法能够提高真阳性率（类似于统计功效），同时将假阳性率（类似于第一类错误）控制在良好水平；其中，组合 $(q^{(1)},\ q^{(2)}) = (0.9,0.5)$具有最理想的操作特征，正如前文第3.2节末尾所述。

4.3 存在协变量分布偏移时的结果 t

接下来，我们考虑存在协变量分布偏移的情景。首先，令 $\mu = 0$ 为 X_3，并令 X_4 适用于两个区域（因此它们完全不涉及协变量偏移）。其次，在区域 r中将协变量 X_s $(s = 1, 2)$ 的参数固定为 $\mu = 0$，从而得到前三阶矩 $\mathbb{E}_r(X_s) = 0$、$\mathbb{E}_r(X_s^2) = 1.61$ 和 $\mathbb{E}_r(X_s^3) = 0$，以及 $\delta_r = 0$。为引入非零平均处理效应，我们在以下两种情景中分别调整 X_1 的参数 μ 以及互补区域的 X_2 参数：(i) 仅 X_1 发生偏移；(ii) X_1 和 X_2均发生偏移。参数 μ 的取值经过选择，使得在所考虑的 κ_r 和 κ_{-r} 组合下，多区域临床试验的总体检验效能均大于90%。回顾一下，对于所有终点，条件平均处理效应函数均涉及 X_1 和 X_2 。我们在附录的表A3中给出了 X_1和 X_2的所得矩，以及连续型终点下互补区域相应的区域平均处理效应；在表A4中给出了二分类与生存终点下 X_1 和 X_2 所得的一阶矩。

图2 显示了区域 r 在情景 (i) 下相对于其互补区域的经验 CP，其中协变量偏移仅发生在 X_1 这两个



(a) 连续型–线性　(b) 连续型—二次　(c) 连续—三次

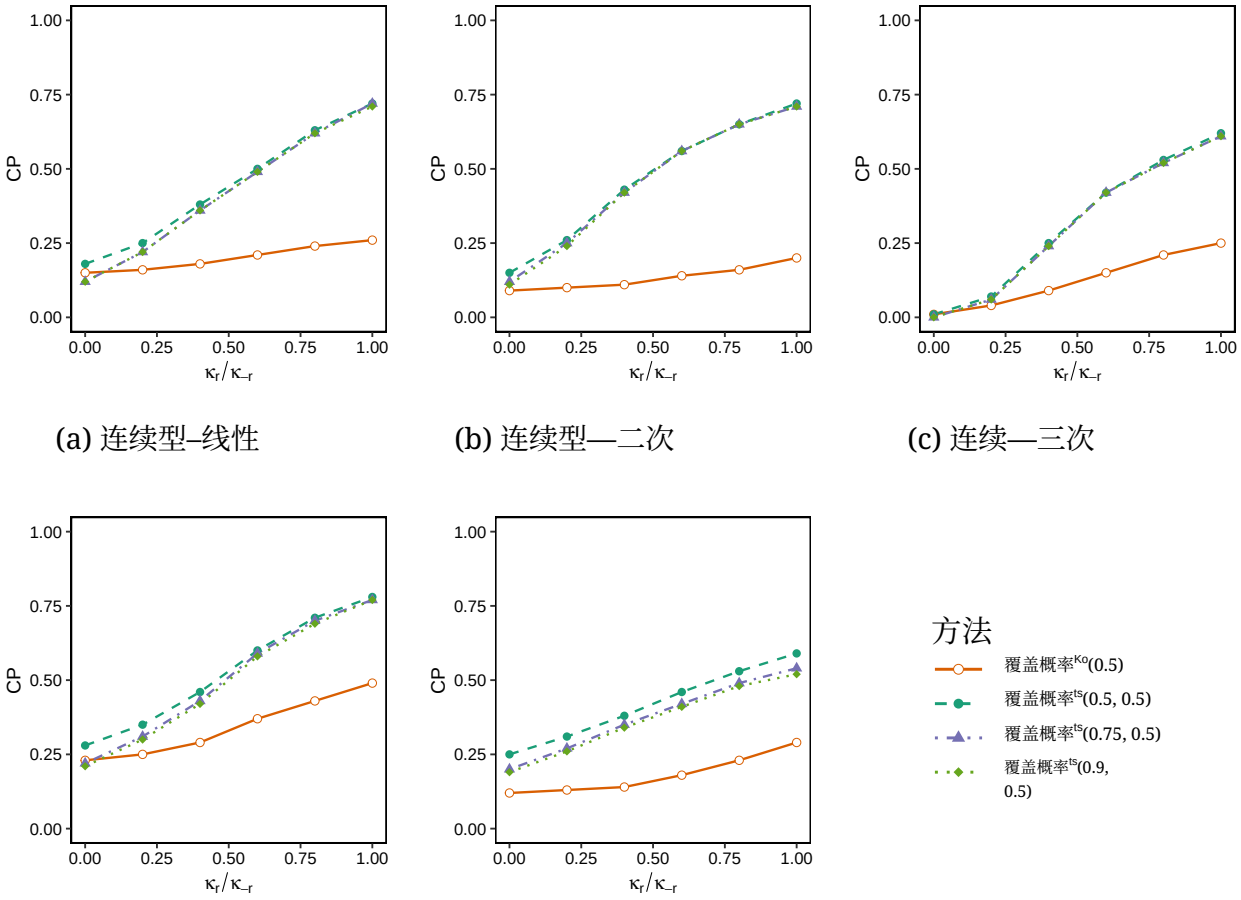
(d) 二元　(e) 生存

图1：在不存在协变量分布偏移的情况下，针对不同终点类型和区域 CATE 的不同差异水平（随 κ_r/κ_{-r} 递减），采用一步法和所提出的两步法得到的区域 r 相对于其互补区域的经验 CP

两步法在所有情况下的表现都显著优于传统一步法。这种性能提升是我们方法的一项关键优势：即使存在协变量偏移，只要区域CATE函数相似（当比率 κ_r/κ_{-r}较大时，例如大于0.6），我们提出的两步法所得到的CP估计值就会远高于一步法。相反，当区域CATE函数不相似（当比率 κ_r/κ_{-r} 较小时，例如小于0.4），两步法的表现往往与一步法相近，得到的CP较低。因此，当区域 r 与其互补区域（$-r$）之间的CATE相似时，我们提出的两步法在处理效应修饰变量发生分布偏移的情况下，能够提高真阳性率（类似于统计功效）；而当CATE不相似时，则能将假阳性率（类似于第一类错误）控制在较低水平，表现出理想的操作特征。

图 3 展示了情景 (ii) 下区域 r 相对于其互补区域的经验 CP，其中 X_1 和 X_2均发生了协变量偏移。研究结果与情景 (i) 类似。也就是说，当各区域的条件平均处理效应不同时，我们提出的两步法与一步法得到的覆盖概率相似且较低；而当各区域的条件平均处理效应相似时，该方法则成功纠正了一致性问题，其覆盖概率远高于一步法。

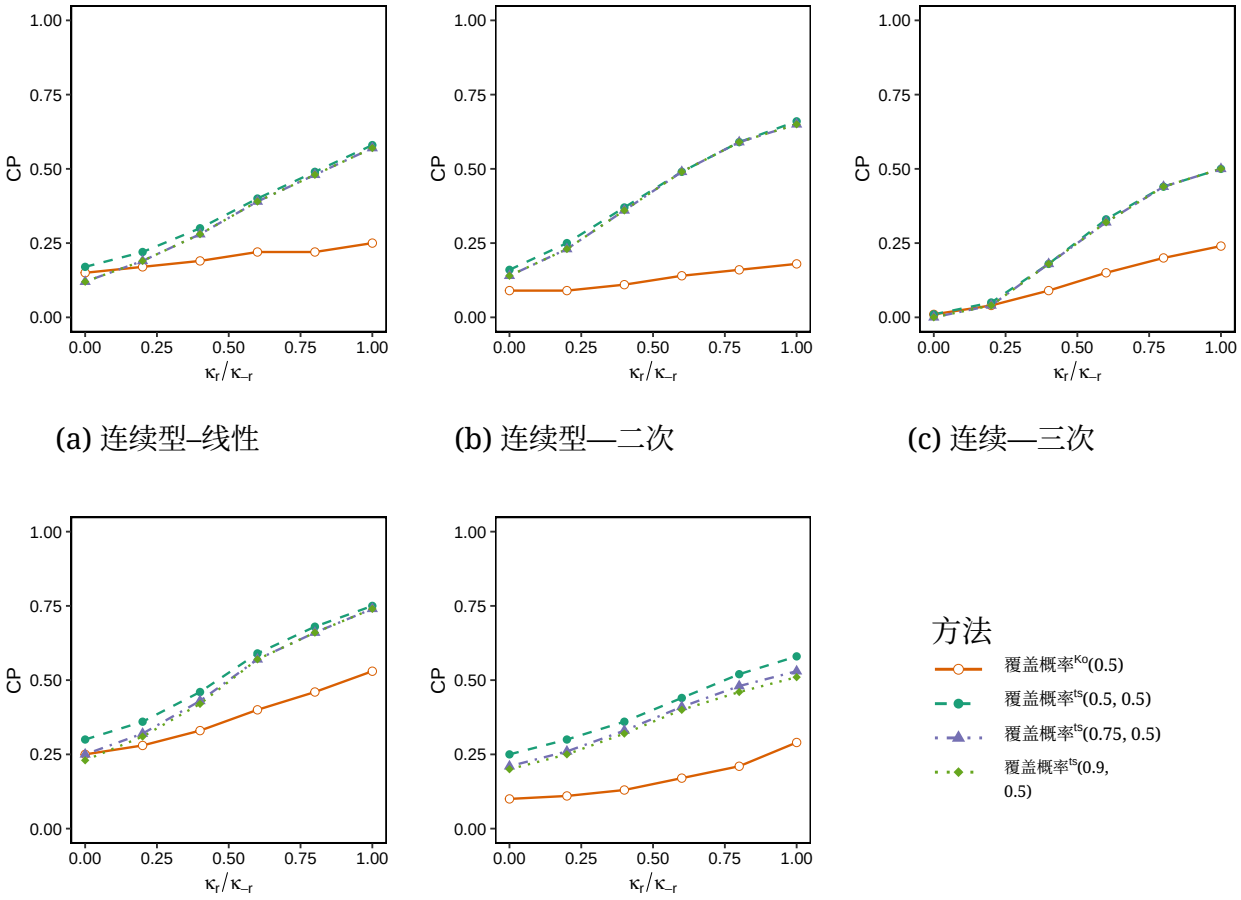
总体而言，我们发现 $\mathrm{CP}^{\mathrm{ts}}(0.5,0.5) \geq \mathrm{CP}^{\mathrm{Ko}}(0.5)$，这与预期一致。为控制条件平均处理效应不同时的假阳性率（类似于第一类错误），并提高条件平均处理效应相似时的真阳性率（类似于统计功效），根据第 4.2 节和第 4.3节模拟结果的评估，我们建议使用 $\mathrm{CP}^{\mathrm{ts}}(0.9,0.5)$。



(a) 连续型–线性　(b) 连续型—二次　(c) 连续—三次

(d) 二元　(e) 生存

图 2：在情景 (i) 下，仅在 X_1 发生协变量偏移时，采用一步法和所提出的两步法得到的区域 r 相对于其互补区域的经验 CP；针对不同终点类型及区域 CATE 差异水平（随 κ_r/κ_{-r} 递减）



(a) 连续型–线性　(b) 连续型—二次　(c) 连续—三次

(d) 二元　(e) 生存

图3：在情景 (ii) 下，通过一步法和所提出的两步法获得的区域 r 相对于其互补区域的经验 CP，其中协变量偏移同时发生在 X_1 和 X_2，并针对不同终点类型及不同程度的区域 CATE 差异（在 κ_r/κ_{-r} 中递减）进行了分析

5 BELIEVE 试验的应用

我们通过 BELIEVE 试验（Cappellini 等人，2020年）说明所提出的两步一致性评估方法的应用。该 III 期多区域临床试验评估了罗特西普对输血依赖型 β-地中海贫血患者的疗效。主要终点为二元响应（响应者或非响应者）；响应者定义为在第 13 至 24 周期间观察到输血负担较基线至少降低 33%，且红细胞单位至少减少 2 个。

共有 224 名患者被分配至罗特西普组，112 名患者被分配至安慰剂组。全球试验结果显示治疗效应具有统计学显著性：罗特西普组中有 21.4% 的患者达到主要终点，而安慰剂组中这一比例为 4.5%。由此得到的优势比（OR）估计值为 5.79（95% CI：2.24，14.97）。

对117例亚洲患者的亚组分析显示效应较小且无统计学意义，罗特西普组和安慰剂组的成功率分别为14.8%和5.6%（比值比：3.06；95%置信区间：0.64,14.63）。然而，在基于亚洲人群一个亚组的事后分析中，该亚组由68例在随机化前每24周基线红细胞输血负担（BTB）为6至15单位的患者组成（代表多区域临床试验总体中较低范围的基线输血负担水平），应答率（21.7%对比4.5%）和估计比值比（5.83）与全球结果高度一致。临床和监管审评表明，较高的基线输血负担水平对应较小的治疗获益，并将基线输血负担视为关键效应修饰变量。这促成了罗特西普在中国输血依赖型 β-地中海贫血患者中获得附条件批准，适用范围限于“基线输血负担 $\leq$ 15单位/24周”。

由于无法获得该试验的个体患者数据，我们模拟数据以重现 BELIEVE 研究中报告的汇总结果。我们假设亚洲患者与非亚洲患者的条件平均处理效应函数相同。我们使用逻辑回归模型生成个体患者数据，并将基线输血负担（取值范围为 $[6, 26]$，如 BELIEVE 试验中所报告）设为效应修饰变量。

达到主要终点 $Y = 1$ 的概率建模如下：

$$\operatorname{logit} \mathbb{P}(Y = 1|\mathrm{BTB}, T) = -15 \cdot \frac{T+1}{2} + 1.3 \cdot \mathrm{BTB} \cdot \frac{T+1}{2} + 2.7,$$

其中， T 是治疗指示变量。选择这种特定的逻辑回归模型参数化方式，是因为由此得到的总体成功率、亚组差异和 OR 与 BELIEVE 试验报告的结果高度吻合。我们在 BTB 中引入亚洲人群与非亚洲人群之间的协变量偏移。具体而言，我们重建了亚洲亚组的数据，其中较大比例（约 42%）的患者处于“高 BTB”范围（15–26 个单位）。相比之下，非亚洲组中处于“高 BTB”范围的患者比例较小（约 18%）。为验证数据模拟过程，我们使用上述逻辑回归模型以及所设定的 BTB 分布偏移，成功复现了 BELIEVE 试验报告的结果。我们使用模拟数据进行了 10，000 次重复，并计算了传统一步法和我们提出的两步法的覆盖概率。结果显示，传统一步法与我们提出的两步一致性评估方法之间存在明显差异。

多区域临床试验的总体统计功效为 99%。一步法 $\mathrm{CP}^{\mathrm{Ko}}(0.5)$ 得出的估计一致性概率为 24%。相比之下，两步法 $\mathrm{CP}^{\mathrm{ts}}(0.5, 0.5)$ 、 $\mathrm{CP}^{\mathrm{ts}}(0.75,0.5)$ 和 $\mathrm{CP}^{\mathrm{ts}}(0.9, 0.5)$ 得出的估计一致性概率分别为 70%、69% 和 69%。这些结果支持了评审意见：不同区域的治疗效应确实可能一致，而传统一步法因区域间 BTB 的偏移未能揭示这一点。更重要的是，在存在协变量偏移的情况下，我们提出的一致性评估两步法能够揭示不同区域间潜在的一致效应，从而得出正确的结论。

6 讨论

多区域临床试验的主要目标是证明所有参与区域的治疗效应具有一致性。本文提出了创新方法来评估区域一致性，并对最初由

提出的日本厚生劳动省(2007)广泛采用的标准进行了扩展。我们的方法一大优势在于，能够纠正因重要处理效应修饰变量发生协变量偏移而导致的不一致性错误结论。我们提出的方法可以轻松扩展至非劣效情形（此时零假设变为 $H_0: \delta + M = 0$ 其中 M 为非劣效界值），方法是在 M 上 δ 施加大小为 Z 的偏移，并分别在 $\hat{\delta}_r$、 $\hat{\delta}_{-r}$和$\hat{\delta}^*_r$中，对式（1）和式（6）应用该偏移。尽管数值研究显示出良好的数值表现，仍有必要开展严格的理论分析，以便为如何选择 $q^{(1)}$ 和 $q^{(2)}$ 作为判定一致性的阈值等问题提供指导。

致谢

本研究由默克公司（MSD）研发（中国）有限公司提供资金支持。作者谨此感谢 MSD 的 William Wang 、 Shuping Jiang 和 Ping Yang 博士在本文准备过程中提出的宝贵意见。

参考文献

Andersen, P.K. 和 Perme, M. P.（2010年）。生存分析中的伪观测值。《医学研究统计方法》， 19(1):71–99。

Armstrong, P. W.（2021年）。扩展替格瑞洛的产品说明书：探究 FDA 的监管内幕。《循环》， 144(8):583–585。

Asano, K., Aoi, Y., Kamada, S., Uyama, Y. 和 Tohkin, M.（2021年）。ICH E17 指南实施的注意事项：借鉴日本以往多区域临床试验的经验。《临床药理学与治疗学》， 109(6):1555–1563。

Breiman, L.（2001）。随机森林。《机器学习》， 45:5–32。

Cappellini, M. D., Viprakasit, V., Taher, A., 等（2018年）。BELIEVE 试验：罗特西普在需要定期输注红细胞（RBC）的成年

β-地中海贫血患者中的一项 III 期随机、双盲、安慰剂对照研究结果。《*Blood*》，132：163。

Cappellini, M. D., Viprakasit, V., Taher, A. T., 等（2020年）。罗特西普用于输血依赖型 β-地中海贫血患者的 III 期试验。《新英格兰医学杂志》，382(13):1219–1231。

Chen, J., Quan, H., Binkowitz, B., Ouyang, S. P., Tanaka, Y., Li, G., Menjoge, S. 和 Ibia, E.（2010年）。评估多区域临床试验中治疗效应的一致性：系统综述。《药物统计学》，9(3):242–253。

Chen, R., Chen, G. 和 Yu, M.（2023a）。利用目标样本汇总信息进行因果推广的熵平衡。《生物统计学》，79(4):3179–3190。

Chen, R., Huling, J. D., Chen, G. 和 Yu, M.（2023b）。通过稳健样本加权促进针对目标总体的个体化治疗规则学习。《生物统计学杂志》，111(1):309–329。

Chen, S., Tian, L., Cai, T., and Yu, M.（2017年）。一种用于亚组识别和比较治疗评分的通用统计框架。《生物统计学》，73(4):1199–1209。

Chen, X., Lu, N., Nair, R., Xu, Y., Kang, C., Huang, Q., Li, N., and Chen, H.（2012）。利用多区域临床试验进行区域审批的决策规则及相关样本量规划。《生物制药统计学杂志》，22(5):1001–1018。

Chen, Y. and Yu, M.（2024年）。具有顺序约束分层的随机化II期选择设计。《生物统计学》，80(1):ujad013。

Gibbons, J., Olkin, I., and Sobel, M.（1999）。《人群选择与排序：一种新的统计方法》。《应用数学经典丛书》。工业与应用数学学会。

Ikeda, K. 和 Bretz, F.（2010年）。多区域试验中的日本患者样本量及占比。《药物统计学》，9(3)：207–216。

Kawai, N.、Chuang-Stein, C.、Komiyama, O. 和 Li, Y.（2008）。一种合理划分多区域试验中各个区域样本量的方法。《药物信息杂志》：*DIJ*/药物信息协会，42(2)：139–147。

Ko, F.-S., Tsou, H.-H., Liu, J.-P. 和 Hsiao, C.-F.（2010年）。多区域试验中特定区域的样本量确定。《生物制药统计学杂志》，20(4)：870–885。

Li, G., Quan, H. 和 Wang, Y.（2024年）。多区域临床试验中的区域一致性评估。《生物制药统计学杂志》，第1–13页。

Liang, M. 和 Yu, M.（2022）。一种用于建模效应修饰的半参数方法。《美国统计学会杂志》，117(538)：752–764。

Mahaffey, K. W., Wojdyla, D. M., Carroll, K., Becker, R. C., Storey, R. F., Angiolillo, D. J., Held, C., Cannon, C. P., James, S., Pieper, K. S., Horrow, J., Harrington, R. A. 和 Wallentin, L.（2011）。血小板抑制与患者结局（PLATO）试验中替格瑞洛与氯吡格雷按地理区域比较。《循环》，124(5)：544–554。

日本厚生劳动省（2007）。《全球临床试验基本原则》。日本厚生劳动省。

Pepe, M. S. 和 Fleming, T. R.（1989）。加权Kaplan-Meier统计量：一类用于删失生存数据的距离检验。《生物统计学》，45(2):497–507。

Quan, H., Li, M., Shih, W. J., Ouyang, S. P., Chen, J., Zhang, J. 和 Zhao, P.（2013）。用于多区域临床试验中治疗效应一致性评估的经验收缩估计量。《医学统计学》，32(10):1691–1706。

Quan, H., Zhao, P., Zhang, J., Roessner, M.和Aizawa, K.（2010年）。基于MHLW指南的多区域试验中日本患者的样本量考量。《药物统计学》，9(2):100–112。

Sargent, D. J.和Goldberg, R. M.（2001年）。多臂筛选试验的灵活设计。《医学统计学》，20:1051–1060。

Sheth, S., Taher, A. T., Coates, T. D., Kattamis, A., and Cappellini, M. D.（2023）。输血依赖型 β-地中海贫血患者罗特西普治疗的管理。《英国血液学杂志》，201(5):824–831。

˜Sugiyama, M., Krauledat, M., and MAˇzller, K.-R.（2007）。通过重要性加权交叉验证适应协变量偏移。《机器学习研究杂志》，8(May):985– 1005。

Tanaka, Y., Li, G., Wang, Y., and Chen, J.（2012）。多区域临床试验中治疗效应的一致性。《生物制药统计学杂志》，22(5):988– 1000。

Teng, Z., Chen, Y.-F., and Chang, M.（2017）。考虑多区域临床试验区域审批时的统一附加要求。《生物制药统计学杂志》，27(6):903–917。

Tsong, Y., Chang, W.-J., Dong, X., 和 Tsou, H.-H. (2012). 多区域临床试验中区域治疗效果的评估。《生物制药统计学杂志》, 22(5):1019–1036.

Tsou, H.-H., Chien, T.-Y., Liu, J.-p., 和 Hsiao, C.-F. (2011). 一种评价桥接研究与多区域试验的一致性方法。《医学统计学》, 30(17):2171–2186.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M. 和 Wei, L.-J.（2014）。超越风险比来量化生存分析中的组间差异。《临床肿瘤学杂志》，32(22):2380–2385。

美国食品药品监督管理局（2024年）。关于如何从肿瘤学多区域临床开发项目中生成临床证据的考量。案卷编号：FDA-2024-D- 3163。

Wallentin、L.、Becker、R. C.、Budaj、A.、Cannon、C. P.、Emanuelsson、H.、Held、C.、Horrow、J.、Husted、S.、James、S.、Katus、H.、Mahaffey、K. W.、Scirica、B. M.、Skene、A.、Steg、P. G.、Storey、R. F. 和 Harrington、R. A.（2009年）。急性冠状动脉综合征患者中替格瑞洛与氯吡格雷的比较。《新英格兰医学杂志》，361（11）：1045–1057。

Wand、M. 和 Jones、M.（1994年）。核平滑。Chapman and Hall/CRC，纽约。

Wu、E. 和 Gagnon-Bartsch、J. A.（2018年）。LOOP估计量：调整随机试验中的协变量。《评估评论》，42（4）：458–488。

Wu, S.-C., Xu, J.-F., Zhang, X.-J., Li, Z.-W., 和 He, J.（2020年）。多区域等效性试验中的区域一致性与样本量考量。《药物统计学》，19(6)：897–908。

Xiong, C., Yu, M., 和 Shao, J.（2017年）。单指标对比函数下的治疗推荐与参数估计。《统计理论及相关领域》，1(2)：171–181。

附录

A.1 限制平均生存时间伪观测值的计算

生存分析中的一个主要挑战是存在右删失数据，即并非所有受试者的感兴趣事件都能被观察到。在基于限制平均生存时间估计条件平均处理效应函数时，必须处理这种删失。一个常见且有效的方法是将删失的事件时间数据转换为伪观测值（Andersen 和 Perme, 2010年），然后将其用于标准回归框架。

伪观测方法会生成一组未删失值，这些值对于每位受试者的真实限制平均生存时间而言是（渐近）无偏的。该过程基于如下刀切法（或“留一法”）统计量。

(i) 基于完整数据估计限制平均生存时间： 首先，使用包含 n 名受试者的完整数据集，计算截至时间 τ 的限制平均生存时间估计值，记为 θ。此估计采用能够处理删失的非参数方法，通常通过对生存函数的Kaplan-Meier估计 $S(u)$ 进行积分来获得。即，

$$\widehat{\theta} = \int_0^{\tau} \widehat{S}(u)du.$$

(ii) 基于留一法数据估计限制平均生存时间： 对于每名受试者 i （从1到 n），通过移除受试者 i 来修改数据集。随后，按照(i)中相同的方法，使用这个包含 $n-1$ 名受试者的缩减数据集计算新的限制平均生存时间估计值 $\theta_{(i)}$。

(iii) 计算伪观测值： 受试者 i 的伪观测值 θ_i^{po} 为：

$$\widehat{\theta}_i^{\text{po}} = n\widehat{\theta} - (n-1)\widehat{\theta}_{(i)}.$$

A.2 二元终点和生存终点的条件平均处理效应函数

对于二元终点，区域 h （$h = r, \quad -r$）的条件平均处理效应函数为

$$\Delta_h(\boldsymbol{X}) = \frac{1}{1+\exp\left[-\left\{-5+5(X_3+X_4)+\kappa_h(X_1+0.5X_2)\right\}\right]} - \frac{1}{1+\exp\left[-\left\{-5+5(X_3+X_4)\right\}\right]}.$$

对于参数为 λ 的指数分布生存终点，在截断点 τ 处的限制平均生存时间为 $(1-e^{-\lambda\tau})/\lambda$。因此，给定第 4.1 节中定义的参数 $\lambda_{h,\,t}$，区域 h 的条件平均处理效应为

$$\Delta_h(\boldsymbol{X}) = \frac{1-\exp\left[-\tau\left\{X_3+X_4+0.1\kappa_h(X_1+0.5X_2)\right\}\right]}{X_3+X_4+0.1\kappa_h(X_1+0.5X_2)} - \frac{1-\exp\left\{-\tau\left(X_3+X_4\right)\right\}}{X_3+X_4}.$$

A.3 第 4.2 节和第4.3 节情景设置的详细信息

表 A1：当考察的 κ_r 和 κ_{-r} 组合下多区域临床试验（通过 t 检验）的统计功效均大于 90% 时，连续型终点在 g 的线性、二次和三次形式下的 X_1 矩（其底层参数 μ 分别为 0.8、1.4 和 0.5）。相关的区域平均处理效应 δ_h （$h = r,\ -r$），其中 $\mu = 0$ 对于 X_2、X_3 和 X_4，列于最后一列。

g	X_1	δ_h
线性	$\mathbb{E}(X_1) = 0.64$	$0.64\kappa_h$
二次	$\mathbb{E}(X_1^2) = 2.45$	$0.84\kappa_h$
三次	$\mathbb{E}(X_1^3) = 1.62$	$1.62\kappa_h$

表 A2：二分类与生存终点的 X_1 一阶矩（以及 $\mu = 0$对于 X_2, X_3 和 X_4），使得在所考虑的 κ_r 和 κ_{-r} 组合下，多区域临床试验（通过 t 检验）均具有超过 90% 的统计功效（相应的区域平均处理效应通过对第A.2节给出的条件平均处理效应进行数值积分获得。）

终点	X_1
二元	$\mathbb{E}(X_1) = 0.408$
生存	$\mathbb{E}(X_1) = -0.791$

表 A3：在连续型终点以及两种协变量偏移情景（即 (i) 仅 X_1 发生偏移，以及 (ii) X_1 和 X_2 均发生偏移）下，区域 $-r$ 的 X_1 和 X_2 在 g 的线性、二次和三次形式下的矩，使得在所考虑的 κ_r 和 κ_{-r} 组合下，多区域临床试验（通过 t 检验）均具有超过 90% 的统计功效。所得区域平均处理效应 δ_{-r} 列于最后一列。

情景	X_1	X_2	δ_{-r}
线性 (i)	$\mathbb{E}_{-r}(X_1) = 0.72$	$\mathbb{E}_{-r}(X_2) = 0$	$0.72\kappa_{-r}$
线性 (ii)	$\mathbb{E}_{-r}(X_1) = 0.49$	$\mathbb{E}_{-r}(X_2) = 0.49$	$0.74\kappa_{-r}$
二次 (i)	$\mathbb{E}_{-r}(X_1^2 - 1.61) = 0.94$	$\mathbb{E}_{-r}(X_2^2 - 1.61) = 0$	$0.94\kappa_{-r}$
二次 (ii)	$\mathbb{E}_{-r}(X_1^2 - 1.61) = 0.64$	$\mathbb{E}_{-r}(X_2^2 - 1.61) = 0.64$	$0.96\kappa_{-r}$
三次 (i)	$\mathbb{E}_{-r}(X_1^3) = 1.96$	$\mathbb{E}_{-r}(X_2^3) = 0$	$1.96\kappa_{-r}$
三次 (ii)	$\mathbb{E}_{-r}(X_1^3) = 1.29$	$\mathbb{E}_{-r}(X_2^3) = 1.29$	$1.94\kappa_{-r}$

表A4：区域 X_1 和 X_2 的 $-r$ 在两种协变量偏移情景下的二分类与生存终点结果：（i）仅偏移 X_1；（ii）同时偏移X_1 和 X_2。在所考虑的 κ_r 和 κ_{-r} 组合下，多区域临床试验（通过 t 检验）的统计功效均大于 90%。（相关的区域平均处理效应通过对第 A.2 节给出的条件平均处理效应进行数值积分得到。）

情景	X_1	X_2
二元 (i)	$\mathbb{E}_{-r}(X_1) = 0.408$	$\mathbb{E}_{-r}(X_2) = 0$
二元（ii）	$\mathbb{E}_{-r}(X_1) = 0.245$	$\mathbb{E}_{-r}(X_2) = 0.164$
生存（i）	$\mathbb{E}_{-r}(X_1) = -0.866$	$\mathbb{E}_{-r}(X_2) = 0$
生存 (ii)	$\mathbb{E}_{-r}(X_1) = -0.486$	$\mathbb{E}_{-r}(X_2) = -0.407$

表A5：在不存在协变量分布偏移的情况下，对于不同终点类型，在区域 CATE 的差异（随 κ_r/κ_{-r} 递减）处于不同水平时，采用一步法和所提出的两步法得到的区域 r 相对于其互补区域的经验 CP

情景	方法	κ_r/κ_{-r}					
		0.0	0.2	0.4	0.6	0.8	1.0
连续型–线性	覆盖概率^{Ko}(0。5)	.17	.29	.44	.57	.67	.75
	覆盖概率^{ts}(0.5, 0.5)	.18	.32	.52	.70	.82	.90
	覆盖概率^{ts}(0.75, 0.5)	.11	.26	.48	.69	.81	.89
	覆盖概率^{ts}(0.9, 0.5)	.10	.25	.48	.68	.81	.89
连续型—二次	覆盖概率^{Ko}(0.5)	.11	.24	.42	.57	.68	.74
	覆盖概率^{ts}(0.5, 0.5)	.11	.27	.51	.71	.84	.90
	覆盖概率^{ts}(0.75, 0.5)	.06	.21	.47	.70	.84	.90
	CP^{ts}(0.9, 0.5)	.05	.20	.47	.70	.83	.89
连续—三次	覆盖概率^{Ko}(0.5)	.02	.17	.40	.55	.66	.72
	覆盖概率^{ts}(0.5, 0.5)	.02	.20	.51	.69	.81	.85
	覆盖概率^{ts}(0.75, 0.5)	.00	.16	.49	.68	.80	.84
	CP^{ts}(0.9, 0.5)	.00	.16	.49	.68	.80	.84
二元	覆盖概率^{Ko}(0.5)	.23	.33	.47	.60	.70	.77
	覆盖概率^{ts}(0.5, 0.5)	.27	.40	.57	.73	.83	.89
	覆盖概率^{ts}(0.75, 0.5)	.21	.35	.54	.70	.82	.88
	CP^{ts}(0.9, 0.5)	.19	.33	.53	.70	.81	.88
生存	覆盖概率^{Ko}(0.5)	.14	.21	.31	.43	.57	.69
	覆盖概率^{ts}(0.5, 0.5)	.21	.32	.44	.59	.71	.81
	覆盖概率^{ts}(0.75, 0.5)	.16	.25	.37	.52	.64	.75
	CP^{ts}(0.9, 0.5)	.14	.23	.35	.49	.61	.71

表 A6：在情景 (i) 下，当仅在 X_1 发生协变量偏移时，对于不同终点类型及不同区域 CATE 的差异水平（随 κ_r/κ_{-r} 递减），采用一步法和所提出的两步法获得的区域 r 相对于其互补区域的经验 CP

情景	方法	κ_r/κ_{-r}					
		0.0	0.2	0.4	0.6	0.8	1.0
连续型–线性	覆盖概率^{Ko}(0。5)	.15	.16	.18	.21	.24	.26
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.18	.25	.38	.50	.63	.72
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.12	.22	.36	.49	.62	.72
	CP$^{\text{ts}}(0.9, 0.5)$	.12	.22	.36	.49	.62	.71
连续型—二次	覆盖概率^{Ko}(0.5)	.09	.10	.11	.14	.16	.20
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.15	.26	.43	.56	.65	.72
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.12	.25	.42	.56	.65	.71
	覆盖概率^{ts}(0.9, 0.5)	.11	.24	.42	.56	.65	.71
连续—三次	覆盖概率^{Ko}(0.5)	.01	.04	.09	.15	.21	.25
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.01	.07	.25	.42	.53	.62
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.00	.06	.24	.42	.52	.61
	CP$^{\text{ts}}(0.9, 0.5)$	.00	.06	.24	.42	.52	.61
二元	覆盖概率^{Ko}(0.5)	.23	.25	.29	.37	.43	.49
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.28	.35	.46	.60	.71	.78
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.22	.31	.43	.59	.70	.77
	CP$^{\text{ts}}(0.9, 0.5)$	.21	.30	.42	.58	.69	.77
生存	覆盖概率^{Ko}(0.5)	.12	.13	.14	.18	.23	.29
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.25	.31	.38	.46	.53	.59
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.20	.27	.35	.42	.49	.54
	CP$^{\text{ts}}(0.9, 0.5)$	.19	.26	.34	.41	.48	.52

表 A7：在情景 (ii) 下，当 X_1 和 X_2 均发生协变量偏移时，对于不同终点类型及不同区域 CATE 差异水平（在 κ_r/κ_{-r} 中递减），采用一步法和所提出的两步法获得的区域 r 相对于其互补区域的经验 CP

情景	方法	κ_r/κ_{-r}					
		0.0	0.2	0.4	0.6	0.8	1.0
连续型–线性	覆盖概率^{Ko} (0.5)	.15	.17	.19	.22	.22	.25
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.17	.22	.30	.40	.49	.58
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.12	.19	.28	.39	.48	.57
	覆盖概率$^{\text{ts}}(0.9, 0.5)$	.12	.19	.28	.39	.48	.57
连续型—二次	覆盖概率^{Ko} (0.5)	.09	.09	.11	.14	.16	.18
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.16	.25	.37	.49	.59	.66
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.14	.23	.36	.49	.59	.65
	覆盖概率$^{\text{ts}}(0.9, 0.5)$	.14	.23	.36	.49	.59	.65
连续—三次	覆盖概率^{Ko} (0.5)	.01	.04	.09	.15	.20	.24
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.01	.05	.18	.33	.44	.50
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.00	.04	.18	.32	.44	.50
	覆盖概率$^{\text{ts}}(0.9, 0.5)$	.00	.04	.18	.32	.44	.50
二元	覆盖概率^{Ko} (0.5)	.25	.28	.33	.40	.46	.53
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.30	.36	.46	.59	.68	.75
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.25	.32	.43	.57	.66	.74
	CP$^{\text{ts}}(0.9, 0.5)$	.23	.31	.42	.57	.66	.74
生存	覆盖概率^{Ko} (0.5)	.10	.11	.13	.17	.21	.29
	覆盖概率$^{\text{ts}}(0.5, 0.5)$	.25	.30	.36	.44	.52	.58
	覆盖概率$^{\text{ts}}(0.75, 0.5)$	.21	.26	.33	.41	.48	.53
	CP$^{\text{ts}}(0.9, 0.5)$	.20	.25	.32	.40	.46	.51