arXiv:2302.01246v2 [stat.AP] 5 Mar 2024

交叉试验中的行为性残留效应与效能考量

交叉试验

Danni Shi 和 Ting Ye*

华盛顿大学生物统计系，美国华盛顿州西雅图，邮编 98195

2024年3月6日

摘要

在不存在残留效应时，交叉试验是一种高效的试验设计。为减小生物学残留效应的影响，通常会设计洗脱期。然而，当洗脱期不符合伦理要求，或无法充分减弱残留效应的影响时，残留效应仍然是一个重要问题。后一种情况可能出现在比较效果研究中，因为其中的残留效应通常不是生物学性的，而是行为性的。本文在潜在结果框架下，研究了存在和不存在残留效应时的交叉设计。我们发现，当残留效应存在且满足符号条件时，基本估计量会低估治疗效应；这不会导致单侧检验的第一类错误膨胀，但会对检验效能产生负面影响。这造成了交叉设计与平行组设计之间的效能权衡。我们推导出了相应条件，在该条件下，交叉设计不会导致第一类错误膨胀，同时其检验效能仍高于平行组设计。我们还针对交叉试验提出了协变量调整方法。最后，我们利用 MTN-034/REACH研究的数据评估了交叉设计和协变量调整的表现。

关键词：因果推断；比较效果研究；协变量调整；效率

*tingye1@uw.edu。此项工作得到了HIV预防试验网络（HPTN）和美国国立卫生研究院（NIH）资助项目的支持：美国国家过敏和传染病研究所（NIAID）5UM1 AI068617。

1 引言

交叉试验是一种纵向研究，患者被随机分配到不同的治疗顺序，以比较不同治疗的效果（Hills和Armitage，1979年；Senn，1994年；Kim等人，2021，第8章）。在最简单的交叉设计中，患者被随机分配到两组：一组在第1期接受治疗1，在第2期接受治疗0；另一组则按相反顺序接受治疗。交叉设计的主要优势是每位患者都可以作为自己的对照，这在很大程度上消除了患者间的变异，使交叉设计比平行组设计更加高效。

交叉设计的一项关键假设是不存在残留效应：也就是说，第1期分配的治疗不会影响第2期的结局。因此，交叉设计最常用于停用后效果便会消失的治疗，以及非吸收性终点。例子包括早期试验（如药代动力学（PK）研究和剂量探索研究）以及针对慢性疾病（如高血压、疼痛和哮喘）的III期研究；综述见 Jones 和 Lewis（1995）。与此同时，通常会在第1期和第2期之间设置洗脱期，以有效降低第1期所接受治疗所造成的生物学残留效应。洗脱期可以是被动的（即洗脱期内不给予治疗），也可以是主动的（即洗脱期内给予治疗，但延迟测量，直至达到稳态）（Senn，2002；Araujo 等，2016）。然而，当洗脱期不适用或不足以充分减轻第1期影响时，残留效应仍是一个亟待关注的问题。例如，MTN-034/REACH研究（Nair等，2023）是一项IIa期HIV预防试验，旨在评估每月达匹韦林阴道环（DVR）和每日口服富马酸替诺福韦二吡呋酯加恩曲他滨（TDF/FTC）在16至21岁青春期女孩和年轻女性中的安全性和依从性。该试验采用两期交叉设计，但不设被动洗脱期，因为不给有感染风险的人群提供有效的HIV预防药物是不合乎伦理的。如果最初使用每月 DVR 的参与者（置入阴道环后，一个月内无需主动采取措施即可保持依从）发现第2期改为每日服药尤其麻烦，那么每月 DVR 在第1期带来的便利和习惯可能会延续至第2期，并影响她们在第2期的依从行为。这类残留效应并非生物学效应，且难以通过洗脱期消除；我们将其称为行为性

残留效应，即治疗通过改变参与者的行为而对后续结局产生的影响。在其他HIV预防试验中，例如TRIO研究（Minnis等，2018），以及更广泛地说，在采用交叉设计的比较效果研究中，这也是一项重要的考量（Hemming等，2020）。

残留效应的存在可能导致治疗效应估计产生偏倚。这促使 Grizzle（1965）提出两阶段程序：首先检验是否存在残留效应，再根据检验结果决定是否在分析中使用第2期数据。Freeman（1989）随后批评了这一两阶段程序，指出其检验效能较低，并可能使后续分析的第一类错误膨胀。另一类方法是对残留效应进行建模（Brown Jr，1980；Laird 等，1992；Jones 和 Donev，1996；Kunert 和 Stufken，2002；Bailey 和 Kunert，2006），但这类方法对模型设定错误较为敏感。

本文采用潜在结果框架（Neyman，1923；Rubin，1974）考察交叉试验中的残留效应，并且只需满足最低限度的统计假设，以更好地理解其影响。在不存在残留效应的理想情形下，我们推导了治疗效应基本估计量的渐近性质，并据此表明，交叉设计通常比平行组设计更有效率。我们的研究结果与文献中已有的效率比较一致（例如，参见 Senn（2002，第9.2.1节）），但结论更为一般，因为我们仅假设不存在残留效应，不作其他假设（如建模假设或数据正态性假设）。随后，我们研究存在残留效应的情形，揭示其对估计偏倚以及检验的第一类错误和检验效能的影响。残留效应为负时，基本估计量会高估治疗效应，从而可能使单侧检验的第一类错误膨胀。相反，残留效应为正时，基本估计量会低估治疗效应；这不会使单侧检验的第一类错误膨胀，但会对检验效能产生不利影响。这导致交叉设计与平行组设计之间存在检验效能权衡。我们还推导出相应条件：在该条件下，交叉设计不会导致第一类错误膨胀，且仍比平行组设计具有更高的检验效能。

本文其余部分安排如下。第2节聚焦于不存在残留效应的交叉试验，介绍交叉试验的潜在结果框架，探讨基本估计量，并比较交叉设计与平行组设计的效率。第3节将结果推广至存在残留效应的情形。第4节提出一种协变量调整方法，并研究其渐近性质。第5节考察交叉设计与

平行组设计之间的效能权衡。第6节介绍我们的方法在REACH研究中的应用，并利用REACH研究的数据考量未来HIV预防试验设计中的检验效能。第7节以讨论作结。

2 无残留效应的交叉试验

2.1 研究设置与假设

考虑一个包含两种处理、两个时期的交叉试验。将一组 n 名受试者随机分配到两种治疗序列中，其中 $A_i = 1$ 表示受试者 i 先接受治疗1，再接受治疗0； $A_i = 0$ 表示相反的顺序。令 $Y_{i1}^{(j)}$ 表示受试者在时期1接受处理 j 时，时期1的潜在结局，其中 $j = 0,\ 1$。令 $Y(jk)\,_{i2}$ 表示受试者在时期1接受处理 j 、在时期2接受处理k时，时期2的潜在结局，其中 $j,k = 0,\ 1$。受试者 i 在时期 t 的观测结局为 Y_{it}。全文中，我们作出一致性假设，将观测结局与潜在结局联系起来：对于 $A_i = 1$, $Y_{i1} = Y(1)\,_{i1}$ 和 $Y_{i2} = Y(10)\,_{i2}$；对于 $A_i = 0, Y_{i1} = Y(0)\,_{i1}$ 和 $Y_{i2} = Y(01)\,_{i2}$。令 $\boldsymbol{X}_i$为受试者 i 的观测基线协变量向量。我们假设 $\left(A_i, \boldsymbol{X}_i, Y(j)\,_{i1}, Y(jk)\,_{i2}, j, k = 0, 1\right), i = 1, \ldots, n$ 相互独立同分布，且具有有限的二阶矩；协方差矩阵Var($\boldsymbol{X}_i$)为正定矩阵。

简单随机化将受试者完全随机地分配到两个治疗序列中。这一假设概括于假设1。

假设1 (简单随机化)。 $A_i \perp \left(\boldsymbol{X}_i, Y_{(j)\,i1}, Y_{(jk)\,i2}\right)$ 对于 $j, k = 0, 1$, 和 $P(A_i = 1) = \pi_1$其中 $0 < \pi_1 < 1$ 已知且 $\pi_0 = 1 - \pi_1$。

假设2是交叉试验中通常采用的关键假设：它指出，时间1的处理对时间2的结局没有直接影响；图1(a)对此进行了示意。许多交叉试验会计划在两个时期之间设置足够长的洗脱期，以使这一假设更为合理。

假设2 (无残留效应)。 对于 $k = 0,1$, $Y_{(0k)\,i2} = Y_{(1k)\,i2} := Y_{(k)\,i2}$ 几乎必然成立。

在假设2下，我们可以简单地令 $Y_{(k)\,i2}, k = 0,1$ 表示时间2的潜在结局。我们关注平均治疗效应 $2^{-1}(\theta_1 + \theta_2)$，其中 $\theta_1 = E\left(Y_{i1}^{(1)} - Y_{i1}^{(0)}\right)$ 和 $\theta_2 = E\left(Y_{i2}^{(1)} - Y^{(0)}{}_{i2}\right)$。请注意，通常会假设治疗效应不随时间变化，因此 $\theta_1 = \theta_2$ （Hills和Armitage，1979年）。然而，如下文定理1所示，研究交叉试验中的治疗效应并不需要假设治疗效应不随时间变化。

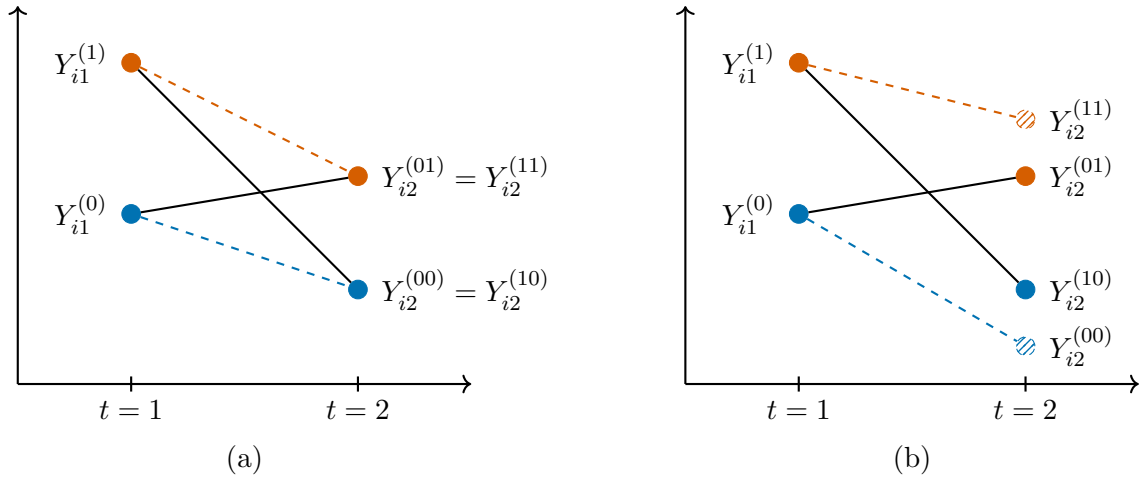

图1：（a）假设2下的六种潜在结局。（b）不满足假设2时的六种潜在结局。

2.2 基本估计量

利用完整交叉周期，我们可以计算每位受试者在治疗1和治疗0下的结局差异。通常使用基本估计量，先计算组别特异性平均结果差异，然后对两个组别的均值取平均（Hills和Armitage，1979年；Senn，1994年；Kim等人，2021，第8章）：

$$\hat{\theta}_{\text{cr}} = \frac{1}{2}\left\{\frac{1}{n_1}\sum_{i=1}^{n} A_i(Y_{i1} - Y_{i2}) + \frac{1}{n_0}\sum_{i=1}^{n}(1 - A_i)(Y_{i2} - Y_{i1})\right\} := \frac{\bar{\Delta}_1 - \bar{\Delta}_0}{2}, \tag{1}$$

其中，$\bar{\Delta}_a = n_a^{-1}\sum_{A_i=a}\Delta_i$ 是具有 $A_i = a$ 的受试者的 Δ_i 的平均值，n_a 是具有 $A_i = a$ 的受试者人数，$a = 0, 1$。定理1总结了基本估计量 θ_{cr} 的统计性质^。

定理1。在假设*1-2*下，

(a) $E(\hat{\theta}_{\text{cr}}) = 2^{-1}(\theta_1 + \theta_2)$，其中 $\theta_t = E(Y(1)_{it} - Y(0)_{it})$ 对于 $t = 1, 2$。*(b)* $\sqrt{n}\{\hat{\theta}_{\text{cr}} - 2^{-1}(\theta_1 + \theta_2)\} \xrightarrow{d} N(0, \sigma_{\text{cr}}^2)$，其中 $\sigma_{\text{cr}}^2 = (4\pi_1)^{-1}Var(Y(1)_{i1} - Y(0)_{i2}) + (4\pi_0)^{-1}Var(Y(0)_{i1} - \bar{Y}_{i2}^{(1)})$。

定理1的证明及所有其他证明均见补充材料。在定理 1(a) 的证明中，我们说明，对于 $A_i = 1$，结果的预期变化等于时间1的平均处理效应减去未接受处理时结果的预期变化，即 $E(Y_{i1} - Y_{i2} \mid A_i = 1) = E(Y_{(1)i1} - Y_{(0)i2}) = E(Y_{(1)i1} - Y_{(0)i1} + Y_{(0)i1} - Y_{(0)i2}) = \theta_1 - \tau$，其中 $\tau = E(Y_{(0)i2} - Y_{(0)i1})$ 表示未接受处理时结果的预期时间变化。同理，

对于 $A_i = 0$，结果的预期变化等于平均处理效应加上未接受处理时结果的预期变化，即 $E(Y_{i2} - Y_{i1} \mid A_i = 0) = E\big(Y_{i2}^{(1)} - Y_{i1}^{(0)}\big) = E\big(Y_{(1)i2} - Y_{(0)i2} + Y_{(0)i2} - Y_{(0)i1}\big) = \theta_2 + \tau$。乍看之下，似乎两组各自的结果平均变化都存在由 τ 导致的偏倚。然而，随机化抵消了非处理因素造成的时间趋势的影响，因此结果的总体平均变化仍是$2^{-1}(\theta_1 + \theta_2)$的无偏估计。此外，在给定样本量 n 时，定理 1(b) 有助于选择使渐近方差最小的最优 π_1。例如，如果 $\mathrm{Var}\big(Y_{(1)i1} - Y_{(0)i2}\big) = \mathrm{Var}\big(Y_{(0)i1} - Y_{i2}^{(1)}\big)$，则最优选择是等比例分配，且 $\pi_1 = \pi_0 = 1/2$。对于统计推断，可通过以下方法一致地估计 σ^2_{cr}

$$\hat{\sigma}^2_{\mathrm{cr}} = \frac{1}{4\pi_1} S^2_{\Delta 1} + \frac{1}{4\pi_0} S^2_{\Delta 0}, \tag{2}$$

其中，$S^2_{\Delta 1}$ 和 $S^2_{\Delta 0}$ 分别是在 $A_i = 1$和 $A_i = 0$处理下受试者的 $\Delta_i = Y_{i1} - Y_{i2}$ 样本方差。

还有一个估计量与基本交叉设计估计量$\hat{}\ \theta_{\mathrm{cr}}$相似：$\hat{\theta}^{\mathrm{alt}}_{\mathrm{cr}} = n^{-1} \sum_{i=1}^{n} \{A_i(Y_{i1} - Y_{i2}) + (1 - A_i)(Y_{i2} - Y_{i1})\}$。在强制执行 $n_1 = n_0$的随机化方案下，这两个估计量$\hat{}\ \theta_{\mathrm{cr}}$ 和$\hat{}\theta^{\mathrm{alt}}_{\mathrm{cr}}$相同；但在其他情况下（例如本文所考虑的简单随机化下），二者并不相同。具体而言，在等比例分配的简单随机化下（$\pi_1 = \pi_0 = 1/2$），$\hat{}\theta^{\mathrm{alt}}_{\mathrm{cr}}$同样是 $2^{-1}(\theta_1 + \theta_2)$的无偏估计量；然而，$\sqrt{n}\big\{\hat{}\theta^{\mathrm{alt}}_{\mathrm{cr}} - 2^{-1}(\theta_1 + \theta_2)\big\}$ 的渐近方差等于 $2^{-1}\mathrm{Var}\big(Y_{(1)i1} - Y_{(0)i2}\big) + 2^{-1}\mathrm{Var}\big(Y_{(0)i1} - Y_{(1)i2}\big) + 4^{-1}(\theta_1 - \theta_2 - 2\tau)^2$，，大于$\hat{}\ \theta_{\mathrm{cr}}$的渐近方差。附加的方差成分 $4^{-1}(\theta_1 - \theta_2 - 2\tau)^2$ 源于两个时间点上时间趋势和处理效应异质性的随机影响。在不等比例分配的简单随机化下，$\hat{}\theta^{\mathrm{alt}}_{\mathrm{cr}}$由于时间效应 τ而存在偏倚。Hills和 Armitage（1979年）也讨论了这一重要问题。因此，本文其余部分不考虑这一替代估计量$\hat{}\theta^{\mathrm{alt}}_{\mathrm{cr}}$。

2.3 交叉设计与平行组设计之间的效率比较

当不存在残留效应时，也就是说，当假设2成立时，众所周知，交叉设计通常比平行组设计效率更高。这是因为每位受试者都可以作为自己的对照，从而使样本量减少一半，而受试者内比较还能进一步消除受试者间变异（Jones and Kenward，2015）。此前的研究（如 Senn（2002，第 9.2.1节））已在不存在残留效应的情况下，并基于各种额外假设（例如建模

假设、时间不变的治疗效应和数据正态性），比较了交叉设计与平行组设计的效率。本节将进行更一般的效率比较，该比较仅假设假设2成立，不作其他假设。

假设计划开展一项随机对照试验，以证明某种试验性治疗的优效性或非劣效性。首先，考虑交叉设计，其零假设H_0： $2^{-1}(\theta_1+\theta_2) = \theta^*$，备择假设为 H_A： $2^{-1}(\theta_1+\theta_2) > \theta^*$，其中$\theta^*$为预先指定值， $\theta^* = 0$ 用于优效性检验， $\theta^* > 0$ 用于非劣效性检验。基于基本估计量的检验统计量为 $T_{\mathrm{cr}} = \sqrt{n}\big(\hat{\theta}_{\mathrm{cr}} - \theta^*\big)/\hat{\sigma}_{\mathrm{cr}}$。根据定理1，在 H_0下， $T_{\mathrm{cr}} \xrightarrow{d} N(0,1)$，因此，当且仅当 $T_{\mathrm{cr}} > z_{1-\alpha}$时，我们拒绝 H_0；其中 α 为显著性水平， $z_{1-\alpha}$ 为标准正态分布的（$1-\alpha$）分位数。在局部备择假设 $\sqrt{n}\big\{2^{-1}(\theta_1+\theta_2) - \theta^*\big\} \to \gamma_{\mathrm{cr}}$下，其中常数为 $\gamma_{\mathrm{cr}} > 0$， T_{cr}的检验效能为 Power $_{\mathrm{cr}} \approx \Phi\left(-z_{1-\alpha} + \gamma_{\mathrm{cr}}/\sigma_{\mathrm{cr}}\right)$, 其中 $\Phi(\cdot)$为标准正态分布的累积分布函数， $\approx$ 表示渐近近似。

现在假设我们只使用时间点1的数据，这实际上就是一种平行组设计。平行组设计中与$\hat{}\ \theta_{\mathrm{cr}}$ 对应的是一个简单的均值差$\hat{}\theta_{\mathrm{pr}} = \frac{1}{n_1}\sum_{i=1}^{n} A_i Y_{i1} - \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)Y_{i1}$。在假设1下，我们有 $E\big(\hat{}\ \theta_{\mathrm{pr}}\big) = \theta_1$和 $\sqrt{n}\big(\hat{}\theta_{\mathrm{pr}} - \theta_1\big) \xrightarrow{d} N(0,\ \sigma^2_{\mathrm{pr}})$，其中 $\sigma^2_{\mathrm{pr}} = \pi_1^{-1}\mathrm{Var}\big(Y_{i1}^{(1)}\big) + \pi_0^{-1}\mathrm{Var}\big(Y_{(0)\,i1}\ \big)$。对于零假设 H_0： $\theta_1 = \theta^*$与备择假设 H_A： $\theta_1 > \theta^*$的检验，检验统计量为 $T_{\mathrm{pr}} = \sqrt{n}\big(\hat{}\theta_{\mathrm{pr}} - \theta^*\big)/\hat{}\sigma_{\mathrm{pr}}$, 其中$\hat{}\sigma^2_{\mathrm{pr}}$ 是 σ^2_{pr}的相合估计量。在局部备择假设$\sqrt{n}(\theta_1 - \theta^*) \to \gamma_{\mathrm{pr}}$下，令常数为 $\gamma_{\mathrm{pr}} > 0$，检验效能 T_{pr} 为

$$\mathrm{Power}_{\mathrm{pr}} \approx \Phi\left(-z_{1-\alpha} + \frac{\gamma_{\mathrm{pr}}}{\sigma_{\mathrm{pr}}}\right). \tag{3}$$

令检验效能$_{\mathrm{cr}}$ = 检验效能$_{\mathrm{pr}} = 1-\beta$，我们可得在两种设计下达到检验效能 $1-\beta$所需的样本量：

$$n_{\mathrm{cr}} = \frac{(z_{1-\alpha} + z_{1-\beta})^2\sigma^2_{\mathrm{cr}}}{\{2^{-1}(\theta_1+\theta_2) - \theta^*\}^2} \text{ and } n_{\mathrm{pr}} = \frac{(z_{1-\alpha} + z_{1-\beta})^2\sigma^2_{\mathrm{pr}}}{(\theta_1 - \theta^*)^2}.$$

例如，Jones and Kenward（2015，第2.4节）曾在时间不变的治疗效应等额外假设下推导出这些样本量公式。当$\theta_1 = \theta_2$时，前述两个零假设相同；要用这两种检验达到相同检验效能所需的样本量之比，也称为 Pitman渐近相对

效率，为

$$\frac{n_{\mathrm{cr}}}{n_{\mathrm{pr}}} = \frac{\sigma^2_{\mathrm{cr}}}{\sigma^2_{\mathrm{pr}}} \cdot \frac{(\theta_1 - \theta^*)^2}{\{2^{-1}(\theta_1 + \theta_2) - \theta^*\}^2}.$$

举例来说，考虑一个简单的情形，其中 $\theta_1 = \theta_2$、$\mathrm{Var}\big(Y_{(j)\,it}\big) = \sigma^2$ for $t = 1, 2$ and $j = 0,1$，且 $\mathrm{Cov}\big(Y_{(j)\,i1}, Y_{(1-j)\,i2}\big) = \rho\sigma^2$，其中 $\rho \in [0, 1)$ 是组内相关系数（ICC）。那么$\sigma^2_{\mathrm{cr}} = 2(1-\rho)\sigma^2$ 且 $\sigma^2_{\mathrm{pr}} = 4\sigma^2$，并且 $n_{\mathrm{cr}}/n_{\mathrm{pr}} = (1-\rho)/2$。这意味着交叉设计仅需 $(1-\rho)/2$，即至多为平行组设计所需样本量的一半；这正是交叉设计相较于平行组设计的关键优势。

3 有残留效应的交叉试验

3.1 设计与假设

当交叉试验中存在残留效应时，时间1的处理可能会影响时间2的结局；图1(b)展示了存在残留效应时六种潜在结局的示意图。在这种情况下，随机化仍能保证假设1成立，但假设2不再成立。

我们提出一种参数化表1中六种潜在结局期望值的方法，该方法直接扩展了第2节中使用的参数化。在表1中，$\tilde{\theta}_2 = E\big(Y^{(11)}_{i2} - Y_{(00)\,i2}\big)$表示时间2的平均处理效应，即比较持续接受治疗1时的潜在结局与持续接受治疗0时的潜在结局；$\tilde{\tau} = E\big(Y_{(00)\,i2} - Y_{(0)\,i1}\big)$表示持续接受治疗0时结局的预期时间变化；$\lambda_0 = E\big(Y^{(10)}_{i2} - Y_{(00)\,i2}\big)$ 和 $\lambda_1 = E\big(Y^{(11)}_{i2} - Y_{(01)\,i2}\big)$ 分别表示两种治疗方案下的预期残留效应。

表1：假设2成立与不成立时六种潜在结局均值的参数化。两行阴影部分对应的是永远不会被观察到的潜在结局。

潜在结局均值	假设2成立时	假设2不成立时
$E(Y^{(0)}_{i1})$	μ	μ
$E(Y^{(1)}_{i1})$	$\mu + \theta_1$	$\mu + \theta_1$
$E(Y^{(00)}_{i2})$	$\mu + \tau$	$\mu + \tilde{\tau}$
$E(Y^{(10)}_{i2})$	$\mu + \tau$	$\mu + \tilde{\tau} + \lambda_0$
$E(Y^{(11)}_{i2})$	$\mu + \tau + \theta_2$	$\mu + \tilde{\tau} + \tilde{\theta}_2$
$E(Y^{(01)}_{i2})$	$\mu + \tau + \theta_2$	$\mu + \tilde{\tau} + \tilde{\theta}_2 - \lambda_1$

3.2 基本估计量

ˆ 定理 2 推导了基本估计量 θ_{cr} （定义见式（1））在不作假设2的情况下的统计性质。

定理 2。在假设*1*下，
(a) $E\big(\hat{}\theta_{\mathrm{cr}}\big)=2^{-1}\big(\theta_1+\tilde{\theta}_2-\lambda_0-\lambda_1\big)$。*(b)* $\sqrt{n}\big\{\hat{}\theta_{\mathrm{cr}}-2^{-1}\big(\theta_1+\tilde{\theta}_2-\lambda_0-\lambda_1\big)\big\}\xrightarrow{d}N\big(0,\tilde{\sigma}_{\mathrm{cr}}^2\big)$, 其中 $\tilde{\sigma}_{\mathrm{cr}}^2=(4\pi_1)^{-1}$方差$\big(Y_{(1)\,i1}-Y_{(10)\,i2}\big)+(4\pi_0)^{-1}$方差$\big(Y_{(01)\,i2}-Y_{(0)\,i1}\big)$。

与定理 1(a) 的证明类似，我们证明 $A_i=1$ 的结果预期变化为$E(Y_{i1}-Y_{i2}\mid A_i=1)=E\big(Y(1)_{\,i1}-Y(10)_{\,i2}\big)=E\big(Y_{(1)\,i1}-Y_{(0)\,i1}+Y_{(0)\,i1}-Y_{(00)\,i2}+Y_{(00)\,i2}-Y_{(10)\,i2}\big)=\theta_1-\tilde{\tau}-\lambda_0$，而 $A_i=0$的结果预期变化为 $E(Y_{i2}-Y_{i1}\mid A_i=0)=E\big(Y(01)_{\,i2}-Y(0)_{\,i1}\big)=E(Y(01)_{\,i2}-Y(11)_{\,i2}+Y(11)_{\,i2}-Y(00)_{\,i2}+Y(00)_{\,i2}-Y(0)_{\,i1})=\tilde{\theta}_2+\tilde{\tau}-\lambda_1$。因此，结果的总体平均变化为 $2^{-1}\big(\theta_1+\tilde{\theta}_2-\lambda_0-\lambda_1\big)$。由此，定理 2(a) 表明，在总体层面无残留效应假设（$E\big(Y(1k)_{\,i2}\big)=E\big(Y(0k)_{\,i2}\big)$ 对于 $k=0,1$）下，ˆ θ_{cr} 仍是治疗效应 $2^{-1}\big(\theta_1+\tilde{\theta}_2\big)$的无偏估计量；该假设弱于假设2中所述的个体层面无残留效应假设。也很容易验证，在假设2下，定理1是定理2的特例。最后请注意，在不满足假设2的情况下，式（2）中定义的 $\hat{}\sigma_{\mathrm{cr}}^2$ 仍是 $\tilde{\sigma}_{\mathrm{cr}}^2$ 的相合估计量。

3.3 第一类错误与检验效能分析

存在残留效应时，首先要问的是，使用 T_{cr} 分析交叉试验是否会导致 I类错误膨胀。为回答这个问题，考虑零假设 $H_0:2^{-1}\big(\theta_1+\tilde{\theta}_2\big)=\theta^*$ 与备择假设 $H_A:2^{-1}\big(\theta_1+\tilde{\theta}_2\big)>\theta^*$，其中 θ^*为预先指定的值。 $T_{\mathrm{cr}}=\sqrt{n}\big(\hat{}\theta_{\mathrm{cr}}-\theta^*\big)/\hat{}\sigma_{\mathrm{cr}}$ 的 I 类错误率为

$$
\begin{aligned}
P_{H_0}(T_{\mathrm{cr}}>z_{1-\alpha})&=P_{H_0}\left(\sqrt{n}\cdot\frac{\hat{\theta}_{\mathrm{cr}}-E\big(\hat{\theta}_{\mathrm{cr}}\big)+E\big(\hat{\theta}_{\mathrm{cr}}\big)-\theta^*}{\hat{\sigma}_{\mathrm{cr}}}>z_{1-\alpha}\right)\\
&\approx\Phi\left(-z_{1-\alpha}-\sqrt{n}\cdot\frac{2^{-1}(\lambda_0+\lambda_1)}{\tilde{\sigma}_{\mathrm{cr}}}\right).
\end{aligned}
$$

当 $\lambda_0+\lambda_1=0$ 时，$\hat{}\theta_{\mathrm{cr}}$ 是治疗效应 $2^{-1}\big(\theta_1+\tilde{\theta}_2\big)$ 的无偏估计量，且 T_{cr} 的第一类错误率为 α。当 $\lambda_0+\lambda_1>0$ 时，$\hat{}\theta_{\mathrm{cr}}$ 低估了治疗效应 $2^{-1}\big(\theta_1+\tilde{\theta}_2\big)$，且 ˆ 第一类错误率小于 α，这意味着该检验是保守的。当 $\lambda_0+\lambda_1<0$ 时， θ_{cr} 高估了治疗效应 $2^{-1}\big(\theta_1+\tilde{\theta}_2\big)$，且第一类错误率大于 α，这意味着

该检验无效。因此，即使存在残留效应，当 $\lambda_0 + \lambda_1 \geq 0$ 时，交叉设计仍能控制单侧检验的第一类错误率。

然而，不会增加 I 类错误率的残留效应可能会对检验效能产生负面影响。具体而言，在存在残留效应且处于局部备择假设$\sqrt{n}\{2^{-1}(\theta_1 + \tilde{\theta}_2) - \theta^*\} \to \gamma_{\text{cr}}$、 $\gamma_{\text{cr}} > 0$为常数的情况下， T_{cr} 的检验效能为

$$\text{Power}_{\text{cr}} \approx \Phi\left(-z_{1-\alpha} + \frac{\gamma_{\text{cr}} - \sqrt{n}2^{-1}(\lambda_0 + \lambda_1)}{\tilde{\sigma}_{\text{cr}}}\right). \tag{4}$$

Senn（1997，表1）简要指出了残留效应对检验效能的影响，但未提供检验效能公式。我们的公式（4）给出了存在残留效应时的一般检验效能公式，其结果与 Senn（1997，表1）中的结果相同（其中 $n = 44\lambda_0 + \lambda_1 = 0$、0.5、1、.....5、 $\tilde{\sigma}_{\text{cr}}^2 = 96$、 $2^{-1}(\theta_1 + \tilde{\theta}_2) = 5$、 $\theta^* = 0$、 $\alpha = 0.025$）。公式（4）还表明，要达到 $1 - \beta$ 检验效能所需的样本量为

$$\tilde{n}_{\text{cr}} = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \tilde{\sigma}_{cr}^2}{\{2^{-1}(\theta_1 + \tilde{\theta}_2 - \lambda_0 - \lambda_1) - \theta^*\}^2}.$$

˜举例来说，考虑一项优效性检验，并假设 $\theta_1 = \theta_2 = \theta_{\text{Alt}} > 0$ ，交叉设计与平行组设计之间的 Pitman渐近相对效率为

$$\frac{\tilde{n}_{\text{cr}}}{n_{\text{pr}}} = \frac{\tilde{\sigma}_{\text{cr}}^2}{\sigma_{\text{pr}}^2} \cdot \left\{1 - \frac{\lambda_0 + \lambda_1}{2\ \theta_{\text{Alt}}}\right\}^{-2}.$$

˜当 $\theta_1 + \theta_2 > \lambda_0 + \lambda_1 \geq 0$ 时，即为控制 I 类错误率，残留效应非负且小于治疗效应时，n_{pr} 与 $\tilde{n}_{\text{cr}}$ 之间的差为

$$\tilde{n}_{\text{cr}} - n_{\text{pr}} = c\left\{\frac{2^{-1}(\lambda_0 + \lambda_1)}{\theta_{\text{Alt}}} - \left(1 - \frac{\tilde{\sigma}_{\text{cr}}}{\sigma_{\text{pr}}}\right)\right\},$$

其中， c 是某个正数。因此，要使交叉设计比平行组设计具有更高的效率，残留效应必须较小。具体而言，与我们在第2节末尾的讨论类似：当 $\tilde{\sigma}_{\text{cr}}/\sigma_{\text{pr}} = \sqrt{(1-\rho)/2}$ 时，我们需要 $\frac{2^{-1}(\lambda_0+\lambda_1)}{\theta_{\text{Alt}}} < 1 - \sqrt{(1-\rho)/2}$。当 $\rho = 0.3, 0.5, 0.7$ 时，我们分别需要 $2^{-1}(\lambda_0 + \lambda_1) > 0$ 且小于 $0.41\theta_{\text{Alt}}, 0.5\theta_{\text{Alt}}, 0.61\theta_{\text{Alt}}$，以避免第一类错误膨胀，并使交叉设计比平行组设计具有更高的检验效能。在许多情形下，这种残留效应较小的条件可能是合理的，因为与治疗效应相比，残留效应通常相对较小

。因此，即使存在残留效应，在许多情况下，交叉设计仍可能比平行组设计具有更高的检验效能。在补充材料的 S2节中，我们讨论如何在 $\lambda_0 + \lambda_1 < 0$ 采用敏感性分析方法时控制第一类错误（Rosenbaum，2020年）。

4 交叉试验的协变量调整

监管机构鼓励在随机对照试验的分析中调整预后基线协变量，因为在几乎不增加未调整估计所需的最低限度统计假设的前提下，这种做法有望显著提高效率（FDA，2023）。这种方法通常会使用结局与协变量之间的工作模型，但其估计目标与采用未调整方法时相同，且推断并不依赖于工作模型的正确设定。

近年来，平行组试验中的协变量调整已得到广泛研究。基于半参数理论（Robins 等人，1994年；Tsiatis，2006年），Tsiatis 等人（2008年）建立了一类用于估计平均处理效应的一致且渐近正态估计量。使用线性工作模型时，采用包含所有处理与协变量交互项的异质协方差分析（ANHECOVA）工作模型进行协变量调整，无论模型是否设定错误，都能保证提高效率（Yang 和 Tsiatis，2001年；Lin，2013；Ye 等人，2022、2023b）。尽管人们广泛建议在交叉试验中进行协变量调整（Metcalfe，2010；Mehrotra，2014；Jemielita 等人，2016年），这些最新研究结果尚未推广至交叉试验。

我们考虑使用以下协变量调整后的异质协方差分析估计量，对随机化前测量的基线协变量向量 $\boldsymbol{X}_i$（即随机化前协变量）进行调整

$$\hat{\theta}_{\mathrm{cr,adj}} = \frac{\{\bar{\Delta}_1 - \hat{\boldsymbol{\beta}}_1^T(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}})\} - \{\bar{\Delta}_0 - \hat{\boldsymbol{\beta}}_0^T(\bar{\boldsymbol{X}}_0 - \bar{\boldsymbol{X}})\}}{2},$$

其中，$\bar{\boldsymbol{X}}$ 是所有 $\boldsymbol{X}_i$的样本均值，$\bar{\boldsymbol{X}}_a$ 是具有 $A_i = a$的受试者的 $\boldsymbol{X}_i$的样本均值，Δ_i 和 $\bar{\Delta}_a$ 定义见式（1），而 $\hat{\boldsymbol{\beta}}_a = \left\{\sum_{i:\ A_i=a}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_a)(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_a)^T\right\}^{-1} \sum_{i:\ A_i=a}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_a)\Delta_i$是通过对具有 $A_i = a$的受试者拟合线性工作模型 $E(\Delta_i \mid A_i = a,\ \boldsymbol{X}_i) = \mu_a + \boldsymbol{\beta}_a^T \boldsymbol{X}_i$ 得到的 $\boldsymbol{\beta}_a$ 的最小二乘估计量。

以下启发式解释说明了为何异质协方差分析不会改变估计目标：即使线性工作模型有误，它通常也能提高效率，但绝不会降低效率。由于随机化使

协变量分布达到平衡， $\bar{\boldsymbol{X}}_a$ 和 $\bar{\boldsymbol{X}}$ 估计的是同一个量，因此，$\hat{\boldsymbol{\beta}}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$ 是零的一个“估计量”。所以， $\bar{\Delta}_a - \hat{\boldsymbol{\beta}}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$ 和 $\bar{\Delta}_a$ 对应于同一个估计目标。此外，由于 $n \to \infty$，无论线性工作模型是否正确，$\hat{\boldsymbol{\beta}}_a$ 都会依概率收敛于 $\boldsymbol{\beta}_a = \mathrm{Var}(\boldsymbol{X}_i)^{-1}\mathrm{Cov}(\boldsymbol{X}_i,\ \Delta_i \mid A_i = a)$。因此， $\bar{\Delta}_a - \hat{\boldsymbol{\beta}}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$ 渐近等价于$\bar{\Delta}_a - \boldsymbol{\beta}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$，其方差为

$$\begin{aligned}\mathrm{Var}\{\bar{\Delta}_a - \boldsymbol{\beta}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})\} &= \mathrm{Var}(\bar{\Delta}_a) + \mathrm{Var}\{\boldsymbol{\beta}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})\} - 2\mathrm{cov}\{\bar{\Delta}_a, \boldsymbol{\beta}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})\} \\ &= \mathrm{Var}(\bar{\Delta}_a) - \mathrm{Var}\{\boldsymbol{\beta}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})\}.\end{aligned}$$

因此， $\bar{\Delta}_a - \hat{\boldsymbol{\beta}}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$ 的渐近方差不大于 $\bar{\Delta}_a$ 的方差。这些结果正式表述于定理3中，且均在渐近意义下成立。

定理3。 在假设*1*下，

(a) $\sqrt{n}\{\hat{\theta}_{\mathrm{cr,adj}} - 2^{-1}(\theta_1 + \tilde{\theta}_2 - \lambda_0 - \lambda_1)\} \xrightarrow{d} N\left(0, \tilde{\sigma}^2_{\mathrm{cr,adj}}\right)$，其中 $\tilde{\sigma}^2_{\mathrm{cr,adj}} = (4\pi_1)^{-1}$方差$\left(Y_{(1)i1} - Y_{(10)i2} - \boldsymbol{\beta}_1^T\boldsymbol{X}_i\right) + (4\pi_0)^{-1}$方差$\left(Y_{(0)i1} - Y_{(01)i2} - \boldsymbol{\beta}_0^T\boldsymbol{X}_i\right) + 4^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T$方差$(\boldsymbol{X})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$，且 $\boldsymbol{\beta}_a =$方差$(\boldsymbol{X}_i)^{-1}$协方差$(\boldsymbol{X}_i, \Delta_i \mid A_i = a)$。

(b) 此外，$\tilde{\sigma}^2_{\mathrm{cr,adj}} - \tilde{\sigma}^2_{\mathrm{cr}} = -(4\pi_1\pi_0)^{-1}(\pi_0\boldsymbol{\beta}_1 + \pi_1\boldsymbol{\beta}_0)^T\mathrm{Var}(\boldsymbol{X})(\pi_0\boldsymbol{\beta}_1 + \pi_1\boldsymbol{\beta}_0) \le 0$。

定理3可通过将 Ye 等人（2023b）中的定理1和推论 1应用于以 Δ_i 作为结局来证明。由定理3(b)可知，$\theta_{\mathrm{cr,adj}}$ 的渐近方差不大于 θ_{cr}的渐近方差，且当且仅当 $\pi_0\boldsymbol{\beta}_1 + \pi_1\boldsymbol{\beta}_0 = 0$时等号成立。例如，当 $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = 0$时，也就是协变量与结局变化 Δ_i不相关时，就会出现这种情况。事实上，Ye 等人（2023b）的定理1给出了更强的结论：在所有形式为 $2^{-1}\left[\{\bar{\Delta}_1 - \boldsymbol{b}_1^T(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}})\} - \{\bar{\Delta}_0 - \boldsymbol{b}_0^T(\bar{\boldsymbol{X}}_0 - \bar{\boldsymbol{X}})\}\right]$的线性调整估计量中，$\theta_{\mathrm{cr,adj}}$ 的渐近方差最小，其中 $\boldsymbol{b}_0, \boldsymbol{b}_1$ 是与 $\boldsymbol{X}_i$维数相同的任意固定或随机向量。$\tilde{\sigma}^2_{\mathrm{cr,adj}}$ 的一个相合估计量为

$$\hat{\sigma}^2_{\mathrm{cr,adj}} = \frac{1}{4\pi_1}S^2_{\Delta 1,\mathrm{adj}} + \frac{1}{4\pi_0}S^2_{\Delta 0,\mathrm{adj}} + \frac{1}{4}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)^T\hat{\Sigma}_X(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0),$$

其中，$S^2_{\Delta a,\mathrm{adj}}$ 是基于 $A_i = a$下受试者的 $Y_{i1} - Y_{i2} - \hat{\boldsymbol{\beta}}_a^T\boldsymbol{X}_i$ 的样本方差，适用于$a = 0,1$； Σ_X 是基于整个样本的 $\boldsymbol{X}_i$ 的样本协方差矩阵。可以很容易地基于协变量调整估计量$\theta_{\mathrm{cr,adj}}$构造一个 Z 检验。根据定理3，从渐近意义上看，该检验的检验效能保证高于未调整的对应项 T_{cr} 。

在本文中， 我们重点关注对随机化前协变量进行调整。正如上文所述

如果协变量与结局变化 Δ_i相关（例如，存在治疗与协变量交互作用时，就可能出现这种情况），那么调整随机化前协变量可以降低 Δ_i 的变异性，并提高渐近效率。第 5 节的模拟研究也表明了这一点。然而，在交叉试验中，除非协变量对结局变化有很强的影响，否则这种调整通常只能带来较小至中等程度的益处。此前的研究探讨了对时期依赖的基线协变量进行调整，即对每个时期给予处理之前测量的协变量进行调整（Kenward 和 Roger，2010；Jones 和 Kenward，2015）。不过，需要特别注意的是，洗脱期必须足够长，以确保时期依赖的基线不受先前处理的影响（即不受残留效应影响）。如果担心存在残留效应，则应使用更复杂的方法来调整时期依赖的协变量（例如，纵向G方法（Hern´an 和 Robins，2020，第 21 章））。探索这些方法将是未来研究的一个方向。

5 功效计算

为比较平行组设计和交叉设计的检验效能，我们考虑两种情形。在情形 I 中，我们考虑以下简单的数据生成过程，可通过公式和模拟计算检验效能：

$$
\begin{aligned}
Y_{i1}^{(0)} &= X_{i1} + X_{i2} + X_{i3} + \epsilon_{i1} \\
Y_{i1}^{(1)} &= \theta_1 + X_{i1} + X_{i2} + X_{i3} + \epsilon_{i2} \\
Y_{i2}^{(10)} &= \tilde{\tau} + \lambda_0 + X_{i1} + X_{i2} + bX_{i3} + \epsilon_{i3} \\
Y_{i2}^{(01)} &= \tilde{\tau} + \tilde{\theta}_2 - \lambda_1 + X_{i1} + X_{i2} + bX_{i3} + \epsilon_{i4}.
\end{aligned}
$$

在情形 II 中，我们考虑一个更复杂的、包含治疗与协变量交互作用的潜在结局数据生成过程：

$$
\begin{aligned}
Y_{i1}^{(0)} &= X_{i1} + X_{i2} + X_{i3} + \epsilon_{i1} \\
Y_{i1}^{(1)} &= X_{i1} + X_{i2} + 2X_{i3} + \theta_1(0.5 + I_{(X_{i2}>0)} + X_{i1}X_{i3}) + \epsilon_{i2} \\
Y_{i2}^{(10)} &= X_{i1} + X_{i2} + X_{i3} + \tilde{\tau}(1 + X_{i1}) + \lambda_0(1 + X_{i2}X_{i3}) + \epsilon_{i3} \\
Y_{i2}^{(01)} &= X_{i1} + X_{i2} + 2X_{i3} + \tilde{\tau}(1 + X_{i1}) + \tilde{\theta}_2(1 + 2X_{i1}X_{i3}) - \lambda_1(1 + X_{i2}X_{i3}) + \epsilon_{i4}.
\end{aligned}
$$

在两种情形下，对于 $j=1$、2、3 以及 $k=1$、2、3、4， X_{ij}、 $\epsilon_{ik}\sim N(0,1)$。治疗组指示变量 A_i 服从伯努利分布，且 $\pi_0=\pi_1=1/2$。对于受试者 i，如果 $A_i=1$，则观测结局为 $Y_{i1}=Y(1)_{i1}$、 $Y_{i2}=Y(10)_{i2}$；如果 $A_i=0$，则观测结局为 $Y_{i1}=Y(0)_{i1}$、 $Y_{i2}=Y(01)_{i2}$。观测数据为（$\boldsymbol{X}_i$、 A_i、 Y_{i1}、 Y_{i2}） $i=1$、.....、 n。对于协变量调整估计量，我们针对 $\boldsymbol{X}_i=(X_{i1}$、 X_{i2}、 $X_{i3})^T$进行调整。我们设定 $n=500$、 $\tilde{\tau}=0.2$、 $\alpha=0.025$、 $\theta^*=0$以及 $b\in\{0,\frac{1}{3}\}$。

在情形 I 中，由于数据生成过程较为简单，我们可以直接计算得出 $\mathrm{Var}(Y(0)_{i1})=\mathrm{Var}(Y(1)_{i1})=4$、$\mathrm{Var}(Y(10)_{i2})=\mathrm{Var}(Y(01)_{i2})=b^2+3$、$\mathrm{Cov}(Y(1)_{i1},Y(10)_{i2})=\mathrm{Cov}(Y(0)_{i1},Y(01)_{i2})=b+2$，因此 $\sigma^2_{\mathrm{pr}}=16$ and $\tilde{\sigma}^2_{\mathrm{cr}}=(1-b)^2+2$。在这种情况下，$\mathrm{Var}(\boldsymbol{X}_i)$ 是一个 3×3 单位矩阵，且$\boldsymbol{\beta}_0=\boldsymbol{\beta}_1=(0,0,1-b)^T$。因此，$\tilde{\sigma}^2_{\mathrm{cr,adj}}=2$。根据 Ye 等人（2023b）的推论 1，也很容易仅利用平行组设计计算协变量调整 ANHECOVA 估计量的渐近方差，得到 $\sigma^2_{\mathrm{pr,adj}}=4$。因此，我们根据公式和模拟计算检验效能。图 2 展示了情形 I 下四种单侧检验$T_{\mathrm{pr}},T_{\mathrm{pr,adj}},T_{\mathrm{cr}},T_{\mathrm{cr,adj}}$的第一类错误率和检验效能，其中 $\theta_1=\theta_2=\theta\in[0,0.5]$ and $\lambda_0=\lambda_1=\lambda\in\{-0.1,0,0.1,0.3\}$是根据公式 (3)、(4) 及其协变量调整对应形式计算的。注意，$T_{\mathrm{pr,adj}}$ 是基于 ANHECOVA 估计量，该估计量的形式与$\theta_{\mathrm{cr,adj}}$ 相同，但将 Δ_i 替换为 Y_{i1}。在此设定中， b 的取值仅影响 T_{cr} 的检验效能，而不影响另外两种检验，因此我们仅展示 T_{cr} 的$b=\frac{1}{3}$。补充材料中给出了通过模拟得到的第一类错误率和检验效能，结果显示其与图 2 中的检验效能一致。

在图 2 中， T_{pr} 和 $T_{\mathrm{pr,\ adj}}$ 的检验效能不受 λ 影响，因此可用于基准比较其他检验的性能。当 $\theta=0$ 时，即在不存在治疗效应的零假设下， T_{pr} 和 $T_{\mathrm{pr,\ adj}}$ 的 I 类错误率等于 $\alpha=0.025$。当 $\lambda<0$时， T_{cr}和 $T_{\mathrm{cr,\ adj}}$ 的 I 类错误率大于 α；当 $\lambda=0$时，等于 α；当 $\lambda>0$时，小于 α。换言之，当 $\lambda\geq0$时， T_{cr} 和 $T_{\mathrm{cr,\ adj}}$ 能够控制 I 类错误率；与此同时，随着 λ 增大， T_{cr} 和$T_{\mathrm{cr,\ adj}}$ 会变得越来越保守。

当 $\theta>0$，即在备择假设下，根据第 3.3 节末尾的计算， T_{cr} 和 T_{pr} 的检验效能比较取决于 $\lambda-\theta\left(1-\sqrt{\frac{2-b}{8}}\right)$ 的符号。具体而言，当 $b=0$时， T_{cr} 的检验效能高于 T_{pr}。因此，在没有残留效应（$\lambda=0$）时，我们发现，与 T_{pr}相比， T_{cr} 的检验效能有显著提升；存在较小的残留效应（$\lambda=0.1$）时，当 $\theta>0.2$，T_{cr} 的检验效能高于 T_{pr}；存在较大的残留效应（$\lambda=0.3$）时，对于 $0<\theta\leq0.5$， T_{pr} 的检验效能高于 T_{cr}。此外，由于较大的 b 会导致较大的组内相关系数，因此 T_{cr}的检验效能也会略有提高。最后，无论模型是否正确，对与结局变化相关的协变量进行调整总能提高检验效能。因此，我们

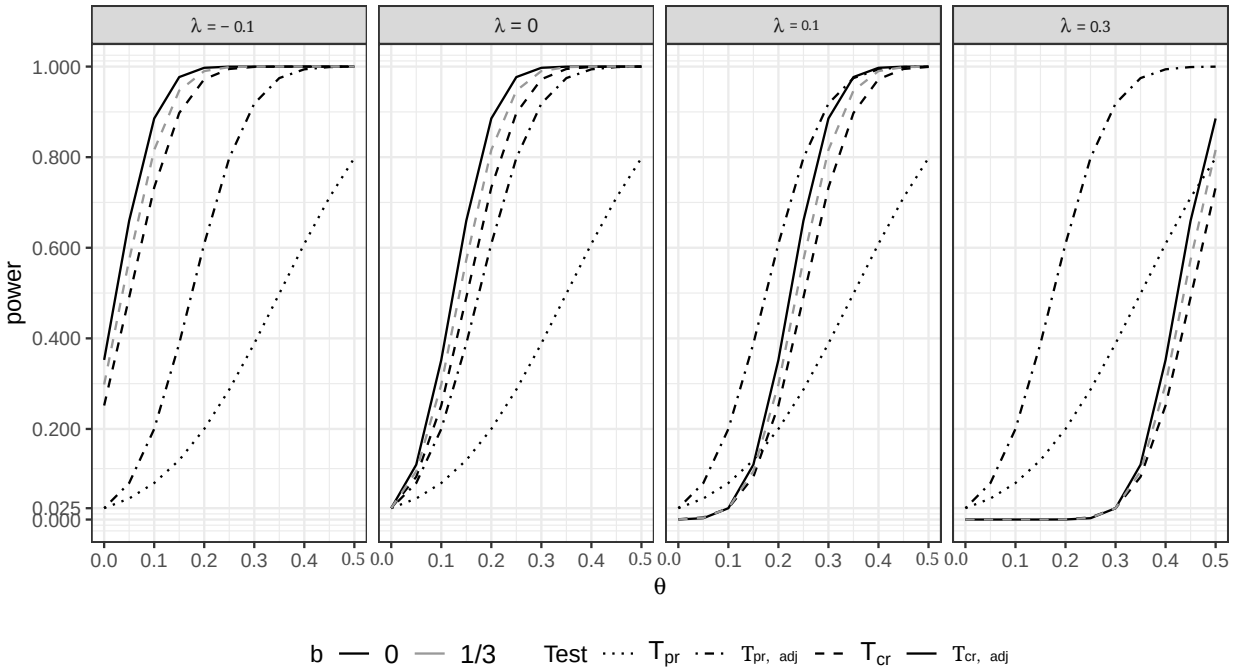

图 2：根据公式计算的情形 I 下四种检验的功效曲线，其中 $\lambda = -0.1$，0，0。1，0。3 和 $b = 0$，1 3 均处于情形 I 下。请注意， T_{pr}、 $T_{\text{pr, adj}}$ 和 $T_{\text{cr, adj}}$ 的检验效能不受 b 的影响。

发现协变量调整检验$T_{\text{cr, adj}}$ 的检验效能始终高于未调整检验 T_{cr}；在某些情况下，检验效能最高可提升 17%。

图3显示了通过10,000次重复模拟获得的情形 II 的经验第一类错误率和检验效能。结果与图3中的结果相似，并且当存在治疗与协变量交互作用时，协变量调整可以提高检验效能。

6 真实数据示例

6.1 REACH研究的应用

在本节中，我们将重新审视第1节所述的 REACH研究（Nair等，2023），并应用我们的方法，研究每月使用DVR与每日口服TDF/FTC之间的依从性差异。

REACH研究共招募247名参与者，并将其随机分配（1:1）至两个产品使用治疗组：一个治疗组使用DVR 6个月，然后改用每日口服TDF/FTC 6个月；另一个治疗组使用每日口服TDF/FTC 6个月，然后改用DVR 6个月。参与者的基线特征汇总于表2。

我们应用该方法比较每月 DVR 和每日口服TDF/FTC的依从性。二元依从性终点定义为第1期和第2期末的高使用率，其中高

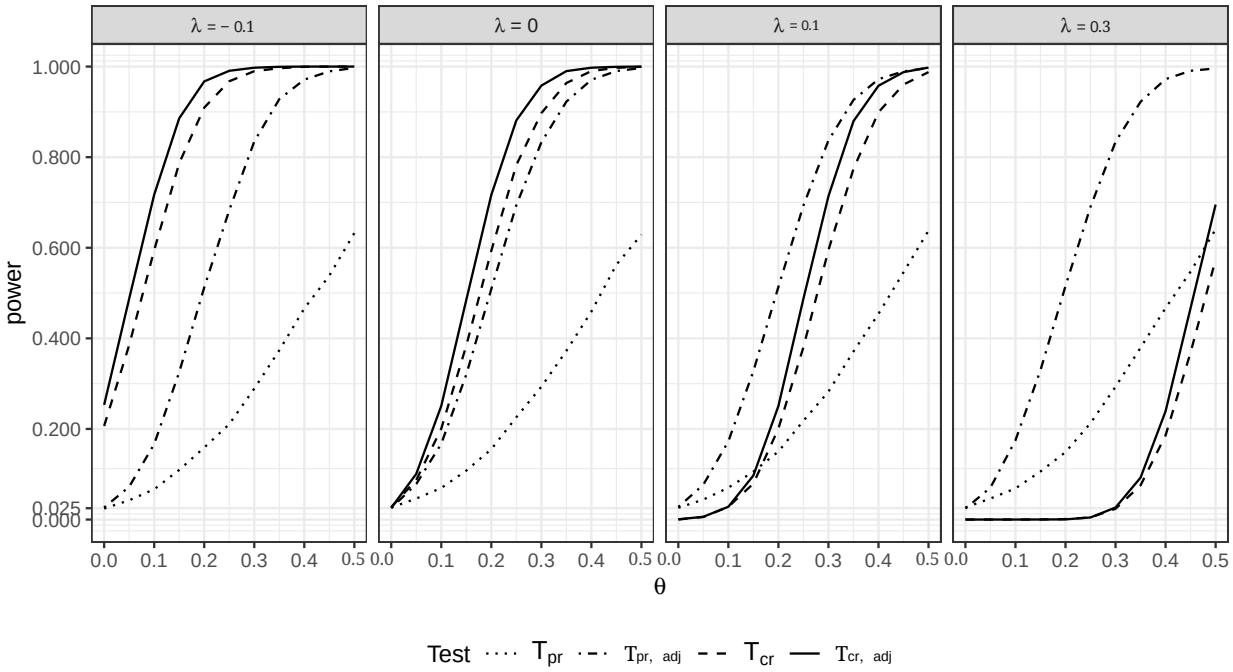

图3：情形 II 下四种检验的功效曲线，由模拟计算得出，其中 $\lambda = -0.1$，0，0。1，0。3。

每日口服TDF/FTC的高使用率定义为替诺福韦二磷酸浓度大于或等于700 fmol/punch（这与前一个月平均每周服用四片或更多片药物有关），每月 DVR 的高使用率则定义为从回收的阴道环中释放出大于或等于4 mg的达匹韦林（前一个月连续使用28天）。文献中广泛采用这些依从性定义（Nair等，2023）。不过，值得注意的是，这些定义要求阴道环每天连续使用，但对口服 TDF/FTC 并无此要求。为进行协变量调整，我们纳入可能与依从性相关的基线协变量，包括年龄、研究地点、对任一产品的偏好、抑郁、酒精使用障碍、性传播感染，以及参与者对HIV感染的担忧程度。我们采用单次插补法处理缺失的基线协变量（Zhao和Ding，2022）。

为讨论潜在残留效应对本试验分析的影响，我们首先指出，对于 DVR 优先组，采用未调整的样本均值，第1期对 DVR 的估计依从率为 55.6%，第2期对 TDF/FTC 的估计依从率为 44.4%；对于 TDF/FTC 优先组，第1期对 TDF/FTC 的估计依从率为 48.0%，第2期对 DVR 的估计依从率为 52.8%。因此，第1期的平均依从性差异为 55.6%-48.0%=7.6%，第2期的平均依从性差异为 52.8%-44.4%= 8.4%。治疗效应在

第1期与第2期之间的这一小幅差异，可能归因于第2期治疗效应较大，和/或存在负向残留效应（λ_1），原因如第1节所述。不过，由于差异很小，即使存在残留效应，也不太可能显著改变结果。

表3列出了平行组和交叉设计的估计值及其协变量调整后的估计值，并同时列出了标准误和95%置信区间。平行组估计值仅使用第1期的数据。所有方法得出的估计值相近，表明与每日口服TDF/FTC相比，DVR的依从性约高出7.7%–8.5%。交叉试验估计量的标准误小于平行组估计量的标准误，说明交叉设计效率更高。调整协变量后，估计值略大，标准误略小。除协变量调整后的交叉设计估计值外，所有95%置信区间均包含0，说明差异不具有统计学显著性。

6.2 使用REACH研究设计假设试验

尽管每日口服TDF/FTC对预防HIV非常有效，但女性对其依从性较低（Celum等，2019年）。双重预防药丸（DPP）是一种每日口服药丸，结合了口服避孕药和TDF/FTC，有望提高女性对每日口服TDF/FTC的依从性（Friedland等，2021年）。本节将利用REACH研究的数据，通过模拟评估在一项假设试验中，交叉设计与平行组设计之间的检验效能权衡。该试验比较单颗每日DPP与每日服用两颗药丸（一颗TDF/FTC和一颗口服避孕药）的依从性。核心假设是，DPP方案可以提高女性对TDF/FTC的依从性。

假设招募基线特征与 REACH研究中参与者相似的青春期女孩和年轻女性参加这项试验，并采用第6.1节所述的相同设计来比较 DPP 与每日两片药。在本例中，我们预计残留效应为非负，主要是因为预期 DPP 会提高第一阶段的依从性，而在第一阶段始终依从的参与者往往会养成每日服用口服药片的习惯，这可能会对她们第二阶段的依从性产生积极影响。根据第3节的结果，在存在非负残留效应的情况下使用交叉设计不会导致第一类错误膨胀。

我们的模拟基于 REACH研究中首先接受 TDF/FTC 的 123 名参与者的数据。在模拟研究中，我们针对不同的治疗效应设定 $\pi_1 = \pi_0 = 0.5$, $\tilde{\tau} = -0.05$, $\theta_1 = \theta_2 = \theta$ ，其中 $\theta \in \{0.05, 0.08, 0.10, 0.15\}$。组内相关系数（即 ρ）估计为

$\rho = 0.39$ 根据 REACH研究中这 123 名参与者的数据。在第 3 节末尾描述的简单情景下，当 $2^{-1}(\lambda_0+\lambda_1)\theta > 1-\sqrt{(1-0.39)/2} \approx 0.447$时，交叉设计的检验效能低于平行设计。因此，我们在此针对不同的残留效应设定 $\lambda_1 = \lambda_0 = \lambda \in \{0, 0.25\theta, 0.45\theta\}$。我们考察四种检验统计量：$T_{\text{pr}}$, $T_{\text{pr,adj}}, T_{\text{cr}}, T_{\text{cr,adj}}$。模拟过程包括以下步骤：

1. 将 REACH研究中基线协变量和第24周的依从性变量分别记为 $(\boldsymbol{X}_i,\ Y(0)_{i1})$，$i = 1\ldots, N$，其中 $N = 123$。$Y(0)_{i1}$ 的平均值为 0.48。使用年龄、研究地点、性传播感染状态、抑郁、酒精使用障碍、担心感染艾滋病毒以及对产品的基线偏好，拟合 TDF/FTC 依从概率的逻辑回归模型。按照第6.1节所述，对所有缺失值进行单次插补。将拟合后的模型记为 $\hat{}\mu_1(\boldsymbol{x}) = \text{expit}(\hat{}\alpha_0 + \hat{}\boldsymbol{\beta}_1^T\boldsymbol{x})$，其中 $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$。

2. 生成 $Y(1)_{i1} \sim$伯努利分布$\left(\text{expit}\left(\hat{}\alpha_1 + \hat{}\boldsymbol{\beta}_1^T\boldsymbol{X}_i\right)\right)$，其中 $\hat{}\alpha_1$满足 $n^{-1}\sum_{i=1}^n \text{expit}(\hat{}\alpha_1 + \hat{}\boldsymbol{\beta}_1^T\boldsymbol{X}_i) - n^{-1}\sum_{i=1}^n \text{expit}\left(\hat{}\alpha_0 + \hat{}\boldsymbol{\beta}_1^T\boldsymbol{X}_i\right) = \theta$。

3. 从 $Y(01)_{i2} \mid Y(0)_{i1} = 1 \sim$伯努利分布生成 $Y(01)_{i2}$，并从 $Y(01)_{i2} \mid Y(0)_{i1} = 0 \sim$伯努利分布生成 $\binom{s_0}{0.48}$，其中选择 s_0（作为 $\tilde{\tau}, \theta, \lambda$ 的函数），以确保 $\text{Corr}(Y(01)_{i2}, Y(0)_{i1}) = \rho$，且 $E\left(Y(01)_{i2}\right) - E\left(Y(0)_{i1}\right) = \tilde{\tau} + \theta - \lambda$；详见补充材料的 S3.3 节。

4. 从 (10) $Y(10)_{i2} \mid Y(1)_{i1} = 1 \sim$伯努利分布生成 $Y(10)_{i2}$，并从 (10) $Y(1)_{i1} = 0 \sim$伯努利分布生成 $\binom{s_1}{0.48+\theta}$，其中选择 s_1（作为 $\tilde{\tau}$、θ、λ 的函数），以确保 $E(Y(10)_{i2}) = 0.48 + \tilde{\tau} + \lambda$，且 $\text{Corr}(Y(10)_{i2}, Y(1)_{i1}) = \rho$。至此，我们已获得 123名个体的协变量和 4 个潜在结局$\left(\boldsymbol{X}_i、\ Y(0)_{i1}、\ Y(1)_{i1}、\ Y(01)_{i2}、\ Y(10)_{i2}\right)$。

5. 从123名个体的协变量和潜在结局中有放回地抽取一个大小为 n 的随机样本。将获得的 n 名受试者随机分配至 $A_i = 0$ 或1，分配概率相等。对于 $A_i = 1$，$Y_{i1} = Y_{i1}^{(1)}$ 和 $Y_{i2} = Y_{i2}^{(10)}$；对于 $A_i = 0$，$Y_{i1} = Y_{i1}^{(0)}$ 和$Y_{i2} = Y_{(01)i2}$。因此，我们得到观测数据（A_i，$\boldsymbol{X}_i$，Y_{i1}，Y_{i2}）$i = 1, \ldots, n$，并计算检验统计量 T_{pr}，$T_{\text{pr, adj}}$，T_{cr}，$T_{\text{cr, adj}}$。

6. 对每个 θ 和 λ，重复步骤（2）至（5）2,000次，据此得到这些检验统计量的经验检验效能。

表4展示了基于2,000次模拟得到的 T_{pr}、$T_{\text{pr、adj}}$、T_{cr}、$T_{\text{cr、adj}}$ 在不同 θ 和 λ 条件下的经验检验效能。样本量 $n = 325$、680、1900的确定方式是使检验效能$_{\text{cr}}$ 接近

80%，此时 $\theta = 0.10$。从表4可以看出，首先，随着 θ增加，经验检验效能也会提高。其次，无论采用平行设计还是交叉设计，协变量调整总能提高检验效能。第三，当 $\lambda = 0$时，交叉设计检验的检验效能明显高于平行组检验。当 $\lambda = 0.25\theta$时，交叉设计检验的检验效能仍然较高，但检验效能差异不那么明显。当 $\lambda = 0.45\theta$时，交叉设计检验的检验效能略低于平行组检验。

7 讨论

交叉设计是一种高效的试验设计，可让参与者充当自己的对照。残留效应，尤其是行为性残留效应，是一个突出问题，因为它可能导致治疗效应的估计产生偏倚。我们采用潜在结果框架和最少的统计假设，考察了两种处理、两个时期的交叉试验中残留效应的影响。我们的结果清晰地刻画了残留效应如何影响估计偏倚，以及检验的第一类错误和检验效能。当残留效应 $\lambda_1 + \lambda_0$ 为负时，基本估计量会高估治疗效应，进而可能使单侧检验的第一类错误率膨胀。相反，当残留效应 $\lambda_1 + \lambda_0$ 为正时，基本估计量会低估治疗效应；这不会使单侧检验的第一类错误率膨胀，但会对检验效能产生负面影响。此外，当 $\lambda_1 + \lambda_0$ 为正、但相较于治疗效应而言较小时，交叉设计的检验效能仍可能高于平行组设计。我们还可以在交叉试验中进一步采用协变量调整，以保证效率提升。本文中的所有方法均可通过 R 中的 RobinCar 软件包实现（Ye 等人，2023a）。

我们结合两项研究讨论研究结果，并证明预期的残留效应可能为非正，也可能为非负：一项是评估每月 DVR 和每日口服 TDF/FTC 的 REACH研究，另一项是假设的未来试验，评估每日服用一片 DPP 和每日服用两片药物。通过这两个示例研究，我们要传达的核心信息是：尽管交叉设计效率较高，但在试验规划和实施过程中，仍需仔细考虑残留效应。

本文主要关注平行组设计与交叉设计之间的检验效能权衡，尤其是在存在残留效应的情况下。然而，这只是各种实际考量之一。首先，交叉试验通常需要更长的随访时间，约为平行组设计的两倍；如果加入洗脱期以减少残留效应，所需时间可能还会

更长（Senn，2002）。更长的随访时间会增加成本和退出风险。其次，交叉设计要求参与者投入更长时间并切换治疗，这可能会令他们望而却步，从而拖慢招募进度。另一方面，在一项研究中有机会尝试多种治疗，在某些情况下可能会让交叉设计对参与者更具吸引力（Diener等，2019）。与此同时，在某些试验中，体验多种治疗对于实现试验目标至关重要。例如，在REACH研究中，参与者必须先体验两种治疗，才能进入选择期，从而研究他们对这两种治疗的偏好。

致谢

我们衷心感谢为本分析提供数据的 MTN-034/REACH研究参与者、研究工作人员和研究者。我们还要感谢匿名审稿人、一位副主编、主编以及 Elizabeth Brown 教授提出的建设性意见。本研究由美国国家过敏症和传染病研究所 [美国国家过敏和传染病研究所 5 UM1 AI068617] 资助。

补充材料

本文所引的第2至第6节网络附录可在 Oxford Academic 的 Biometrics 网站查阅。复现模拟和应用结果所需的代码可在补充材料中找到。

数据可用性

支持本文研究结果的数据可向杀微生物剂试验网络获取。这些数据的获取受到限制，且本研究使用这些数据需遵守许可规定。经杀微生物剂试验网络许可，可向作者获取数据。

参考文献

Araujo, A., Julious, S., 和 Senn, S.（2016年）。理解单病例试验组中的变异。*PLoSONE*, 11(12):e0167167。

Bailey, R. 和 Kunert, J.（2006年）。关于携带效应与直接效应成比例时的最优交叉设计。*Biometrika*，93(3)：613–625。

Brown Jr, B. W.（1980）。临床试验中的交叉实验。*Biometrics*，36：69–79。Celum, C. L., Delany-Moretlwe, S., Baeten, J. M., Straten, A., Hosek, S., Bukusi, E. A. 等（2019）。非洲青春期女孩和年轻女性的 HIV 暴露前预防：从有效性试验到实施。《国际艾滋病学会杂志》，22(S4)：e25298。Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. 和 Wynder, E.（1959）。吸烟与肺癌。《美国国家癌症研究所杂志》，22：173–203。

Diener, H.-C., Tassorelli, C., Dodick, D. W., Silberstein, S. D., Lipton, R. B., Ashina, M., Becker, W. J., Ferrari, M. D., Goadsby, P. J., Pozo-Rosich, P., Wang, S.-J., 和 Mandrekar, J.（2019）。《国际头痛学会关于成年人偏头痛发作急性治疗对照试验的指南：第四版》。*Cephalalgia*，39(6)：687–710。

FDA（2023）。药物和生物制品随机临床试验中协变量的调整。行业指南。药品评价与研究中心和生物制品评价与研究中心，食品药品监督管理局（FDA），美国卫生与公众服务部。2023年5月。

Freeman, P. (1989). 两治疗两周期交叉试验两阶段分析的表现。《医学统计学》, 8(12):1421–1432. Friedland, B. A., Mathur, S., and Haddad, L. B. (2021). 双重预防药丸的前景：开发与引入框架。《生殖健康前沿》, 3(682689). Grizzle, J. E. (1965). 两周期交叉设计及其在临床试验中的应用。*Biometrics*, 21:467–480. Hemming, K., Taljaard, M., Weijer, C., and Forbes, A. B. (2020). 在比较效果研究中使用多周期整群随机交叉试验设计。*BMJ*, 371. Hern´an, M. and Robins, J. (2020). 《因果推断：假设》. 博卡拉顿：Chapman & Hall/CRC. Hills, M. and Armitage, P. (1979). 两周期交叉临床试验。《英国临床药理学杂志》, 8(1):7. Jemielita, T., Putt, M., and Mehrotra, D. (2016). 通过基线值的线性组合提高交叉设计的检验效能。《医学统计学》, 35(30):5625–5641。

Jones, B. and Donev, A. (1996). 交叉试验的建模与设计。《医学统计学》, 15(13):1435–1446. Jones, B. and Kenward, M. G. (2015). 交叉试验的设计与分析. CRC出版社. Jones, B. and Lewis, J. (1995). III期交叉试验的理由。《医学统计学》, 14(9):1025–1038. Kenward, M. G. and Roger, J. H. (2010). 基线协变量在交叉研究中的应用。生物统计学, 11(1):1–17. Kim, K., Bretz, F., Cheung, Y. K. K., and Hampson, L. V. (2021). 随机对照试验统计方法手册. CRC出版社. Kunert, J. and Stufken, J. (2002). 在包含自身及混合残留效应的模型中实现最优交叉设计。美国统计学会会刊, 97(459):898–906. Laird, N. M., Skinner, J., and Kenward, M. (1992). 对存在残留效应的两周期交叉设计进行分析。《医学统计学》, 11(14-15):1967–1979. Lin, W. (2013). 关于实验数据回归调整的不可知论笔记：重新审视Freedman的批评。应用统计年刊, 7(1):295–318. Mehrotra, D. V. (2014). 针对具有基线测量的交叉试验的推荐分析方法。药物统计学, 13(6):376–387. Metcalfe, C. (2010). 对具有基线测量的交叉试验进行分析。《医学统计学》, 29(30):3211–3218. Minnis, A. M., Roberts, S. T., Agot, K., Weinrib, R., Ahmed, K., Manenzhe, K., and et al. (2018). 肯尼亚和南非一项随机交叉研究中，年轻女性对三种用于预防HIV和妊娠的安慰剂多用途预防技术的评价。艾滋病与行为, 22(8):2662–2673. Nair, G., Celum, C., Szydlo, D., Brown, E. R., Akello, C. A., Nakalega, R., and et al. (2023). 非洲青春期女孩和年轻女性对每月使用达匹韦林阴道环或口服恩曲他滨加富马酸替诺福韦二吡呋酯进行HIV暴露前预防的依从性、安全性及选择：一项随机、开放标签交叉试验。《柳叶刀·艾滋病》。

Neyman, J. (1923). 概率论在农业实验中的应用。原理论文。第9节。《统计科学》，5(4):465–472。Dorota M. Dabrowska 和 Terence P. Speed 译（1990）。Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). 当某些回归变量并非总能观测到时回归系数的估计。美国统计学会会刊，89(427):846–866。Rosenbaum, P. R. (2020). 《观察性研究设计》（第*2*版）。施普林格。Rubin, D. B. (1974). 随机研究和非随机研究中治疗因果效应的估计。《教育心理学杂志》，6(5):688–701。Senn, S. (1994). AB/BA交叉设计：过去、现在与未来？《医学研究中的统计方法》，3(4):303–324。Senn, S. (2002). 《临床研究中的交叉试验》。约翰·威立父子有限公司。Senn, S. J. (1997). 致编辑的信：B. J. Jones 和 J. Lewis 在《医学统计学》14卷（1995年）1025–1038页发表的《支持III期交叉试验》一文。《医学统计学》，16(17):2021–22。Tsiatis, A. A. (2006). 半参数理论与缺失数据。Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). 随机临床试验中双样本治疗比较的协变量调整：一种有原则且灵活的方法。《医学统计学》，27(23):4658–4677。Yang, L. and Tsiatis, A. A. (2001). 前测–后测试验中治疗效应估计量的效率研究。《美国统计学家》，55(4):314–321。Ye, T., Bannick, M., Yi, Y., and Bian, F. (2023a). RobinCar：针对协变量自适应随机化的稳健估计与推断，https://github.com/tye27/RobinCar（访问于2023年3月20日）。Ye, T., Shao, J., Yi, Y., and Zhao, Q. (2023b). 改进随机临床试验分析中协变量调整实践。美国统计学会会刊，页码1–13。Ye, T., Yi, Y., and Shao, J. (2022). 最小化法及其他协变量自适应随机化方法下平均处理效应的推断。*Biometrika*，109(1):33–47。

Yi, Y., Zhang, Y., Du, Y., and Ye, T. (2022). 临床试验中使用内部和外部对照两次检验治疗效应。*arXiv*预印本 *arXiv:2203.04194*。

Zhao, A. 和 Ding, P.（2022）。调整还是不调整？在缺失协变量的随机实验中估计平均处理效应。美国统计学会会刊，第 1–11 页。

表2：REACH研究中按随机交叉序列分列的基线特征。数据以中位数（四分位距）或 n (%)表示。

	先使用DVR ($n = 124$)	先使用TDF/FTC ($n = 123$)	总体 (n=247)
年龄（岁）			
中位数（四分位距）	18（17-19）	18（17-19）	18（17-19）
研究地点			
南非 - 开普敦	30 (24.2%)	30 (24.4%)	60 (24.3%)
南非 - 约翰内斯堡	34 (27.4%)	33（26.8%）	67（27.1%）
乌干达 - 坎帕拉	30（24.2%）	30（24.4%）	60（24.3%）
津巴布韦 - 哈拉雷	30 (24.2%)	30 (24.4%)	60 (24.3%)
婚姻状况			
未婚	106 (85.5%)	108 (87.8%)	214 (86.6%)
已婚/同居	16 (12.9%)	14 (11.4%)	30 (12.1%)
分居/离异	2 (1.6%)	1 (0.8%)	3 (1.2%)
最高受教育程度			
小学	15 (12.3%)	18 (14.6%)	33 (13.5%)
中学	92 (75.4%)	97（78.9%）	189（77.1%）
高等教育	15（12.3%）	8（6.5%）	23（9.4%）
有自己的收入	22（17.7%）	31（25.2%）	53（21.5%）
曾经怀孕	52（41.9%）	47 (38.2%)	99 (40.1%)
目前有性伴侣	111 (89.5%)	108 (87.8%)	219 (88.7%)
担心感染艾滋病毒	86（39.4%）	82（66.7%）	168（68.0%）
确诊性传播感染	40（32.3%）	47（38.2%）	87（35.2%）
梅毒	3（2.4%）	3（2.4%）	6（2.4%）
滴虫病	8 (6.5%)	5 (4.1%)	13 (5.3%)
淋病	13 (10.5%)	8 (6.5%)	21 (8.5%)
衣原体感染	30 (24.2%)	41 (33.3%)	71 (28.7%)
CES抑郁量表			
≥ 12 (表明存在抑郁)	28（22.6%）	38（30.9%）	66（26.7%）
< 12	89（71.8%）	78 (63.4%)	167 (67.6%)
缺失	7 (5.6%)	7 (5.7%)	14 (5.7%)
酒精使用障碍（AUDIT-C评分）			
≥ 3 （表明存在酒精使用障碍）	50（40.3%）	47（38.2%）	97（39.3%）
< 3	74（59.7%）	75 (61.0%)	149 (60.3%)
缺失	0 (0.0%)	1 (0.8%)	1 (0.4%)
产品偏好			
DVR	51 (41.1%)	43 (35.0%)	94 (38.1%)
口服 TDF/FTC	43 (34.7%)	57 (46.3%)	100 (40.5%)
偏好相同	29 (23.4%)	22 (17.9%)	51 (20.6%)
未作答	1 (0.8%)	1 (0.8%)	2 (0.8%)

四分位距=四分位距；性传播感染=性传播感染；流行病学研究中心=流行病学研究中心；酒精使用障碍识别测试-简版= 酒精使用障碍识别测试-简版。

表3：与每日口服TDF/FTC相比，DVR对依从性的平均处理效应的协变量调整和未调整平行设计及交叉设计估计量，以及标准误和95%置信区间。

Type	Mean	SE	95%置信区间
平行设计，未调整	0.077	0.064	(−0.048, 0.202)
平行设计，已调整	0.082	0.062	(−0.040, 0.204)
交叉，未调整	0.081	0.043	(−0.003, 0.165)
交叉，已调整	0.085	0.042	(0.003, 0.168)

表4：使用 REACH研究数据进行假设试验的 T_{pr}，$T_{\text{pr, adj}}$，T_{cr}，$T_{\text{cr, adj}}$ 的经验检验效能。

n	λ	θ	检验效能$_{\text{pr}}$	检验效能$_{\text{pr, adj}}$	检验效能$_{\text{cr}}$	检验效能$_{\text{cr, adj}}$
325	0	0.05	0.214	0.236	0.423	0.446
		0.08	0.342	0.379	0.670	0.681
		0.10	0.439	0.484	**0.797**	0.818
		0.15	0.738	0.770	0.973	0.978
680	0.25θ	0.05	0.324	0.354	0.452	0.463
		0.08	0.534	0.570	0.680	0.6701
		0.10	0.679	0.720	**0.799**	0.808
		0.15	0.905	0.923	0.959	0.961
1900	0.45θ	0.05	0.558	0.579	0.523	0.528
		0.08	0.778	0.789	0.684	0.688
		0.10	0.876	0.889	**0.791**	0.797
		0.15	0.988	0.989	0.929	0.935

补充材料

S1 额外功效比较

本节通过 10,000 次重复模拟得到的检验效能，来检验第 5 节公式计算的检验效能。图 S4 只是在图 2 的基础上叠加了模拟检验效能结果；我们发现，公式计算的检验效能与模拟得到的检验效能吻合良好。

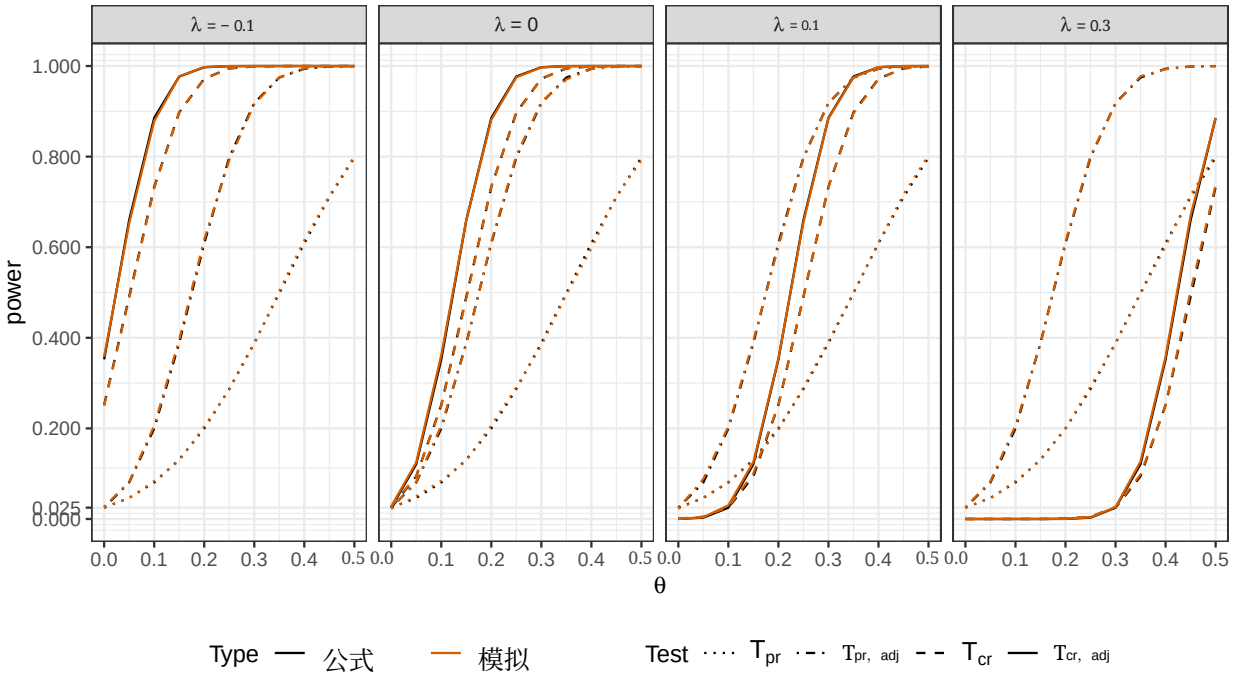

图 S4：当 $\lambda = -0.1$，0，0。1，0。3 和 $b = 0$ 时，情形 I 下四种检验的功效曲线，分别由公式和模拟（重复 10,000 次）计算得到。两种方法得到的功效曲线彼此吻合。

S2 当 $\lambda_0 + \lambda_1 < 0$ 时控制第一类错误

与 Yi 等人（2022）类似，我们考虑一个敏感性参数 $\Lambda < 0$，它为偏倚设定了界限，即$_{1\ 2}($ $\lambda_0 + \lambda_1) \geq \Lambda$。那么拒绝域

$$\frac{\sqrt{n}\left(\hat{\theta}_{\mathrm{cr}} - \theta^* + \Lambda\right)}{\hat{\sigma}_{\mathrm{cr}}} > z_{1-\alpha}$$

可将第一类错误控制在水平 α。

实施敏感性分析时，研究者无需指定敏感性参数的取值 Λ。敏感性参数的结果可以用“临界点”来概括，即 Λ 所需达到的幅度，使得零假设不再能够被拒绝（Cornfield et al., 1959; Rosenbaum, 2020）。如果认为 Λ 的这一取值不可信，那么我们仍有证据拒绝零假设。

S3 技术证明

S3.1 定理1的证明

证明。 (a) 令 $\tau = E\left(Y_{i2}^{(0)} - Y_{i1}^{(0)}\right)$ 表示未接受处理时的预期时间趋势。那么

$$
\begin{aligned}
&E(\hat{\theta}_{\mathrm{cr}}) \\
&= \frac{1}{2}E\left\{\frac{1}{n_1}\sum_{i=1}^{n} A_i(Y_{i1} - Y_{i2}) + \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)(Y_{i2} - Y_{i1})\right\} \\
&= \frac{1}{2}E\left\{\frac{1}{n_1}\sum_{i=1}^{n} A_i\left(Y_{i1}^{(1)} - Y_{i2}^{(0)}\right) + \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)\left(Y_{i2}^{(1)} - Y_{i1}^{(0)}\right)\right\} \\
&= \frac{1}{2}E\left\{\frac{1}{n_1}\sum_{i=1}^{n} A_i\left(Y_{i1}^{(1)} - Y_{i1}^{(0)} + Y_{i1}^{(0)} - Y_{i2}^{(0)}\right) + \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)\left(Y_{i2}^{(1)} - Y_{i2}^{(0)} + Y_{i2}^{(0)} - Y_{i1}^{(0)}\right)\right\} \\
&= \frac{1}{2}(\theta_1 - \tau) + \frac{1}{2}(\theta_2 + \tau) = \frac{\theta_1 + \theta_2}{2}.
\end{aligned}
$$

(b) 根据中心极限定理，渐近正态性显而易见。接下来，我们推导 $\sqrt{n}(\hat{\theta}_{\mathrm{cr}} - \frac{\theta_1+\theta_2}{2})$ 的渐近方差。注意，

$$
\hat{\theta}_{\mathrm{cr}} - \frac{\theta_1 + \theta_2}{2} = \underbrace{\frac{1}{2}\left\{\frac{1}{n_1}\sum_{i=1}^{n} A_i\left(Y_{i1}^{(1)} - Y_{i2}^{(10)} - \theta_1 + \tau\right) + \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)\left(Y_{i2}^{(01)} - Y_{i1}^{(0)} - \theta_2 - \tau\right)\right\}}_{M_1}
$$

我们可以证明

$$
\sqrt{n}M_1 \mid A_1, \ldots, A_n \xrightarrow{d} N\left(0, \frac{\mathrm{Var}\left(Y_{i1}^{(1)} - Y_{i2}^{(0)}\right)}{4\pi_1} + \frac{\mathrm{Var}\left(Y_{i2}^{(1)} - Y_{i1}^{(0)}\right)}{4\pi_0}\right).
$$

根据有界收敛定理，该结果在无条件情况下仍然成立，即，

$$\sqrt{n}\left(\hat{\theta}_{\mathrm{cr}}-\frac{1}{2}(\theta_1+\theta_2)\right)\xrightarrow{d}N\left(0,\frac{\mathrm{Var}\left(Y_{i1}^{(1)}-Y_{i2}^{(0)}\right)}{4\pi_1}+\frac{\mathrm{Var}\left(Y_{i2}^{(1)}-Y_{i1}^{(0)}\right)}{4\pi_0}\right).$$

□

S3.2 定理 2 的证明

(a) 我们按如下方式计算期望:

$$\begin{aligned}E(\hat{\theta}_{\mathrm{cr}})&=\frac{1}{2}E\left\{\frac{1}{n_1}\sum_{i=1}^{n}A_i\left(Y_{i1}^{(1)}-Y_{i2}^{(10)}\right)+\frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)\left(Y_{i2}^{(01)}-Y_{i1}^{(0)}\right)\right\}\\&=\frac{1}{2}E\Big\{\frac{1}{n_1}\sum_{i=1}^{n}A_i\left(Y_{i1}^{(1)}-Y_{i1}^{(0)}+Y_{i1}^{(0)}-Y_{i2}^{(00)}+Y_{i2}^{(00)}-Y_{i2}^{(10)}\right)\\&\quad+\frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)\left(Y_{i2}^{(01)}-Y_{i2}^{(11)}+Y_{i2}^{(11)}-Y_{i2}^{(00)}+Y_{i2}^{(00)}-Y_{i1}^{(0)}\right)\\&=\frac{1}{2}(\theta_1-\tilde{\tau}-\lambda_0)+\frac{1}{2}\left(\tilde{\theta}_2-\lambda_1+\tilde{\tau}\right)=\frac{1}{2}\left(\theta_1+\tilde{\theta}_2-\lambda_0-\lambda_1\right).\end{aligned}$$

(b) 根据中心极限定理，渐近正态性显而易见。接下来，我们推导 $\sqrt{n}(\hat{\theta}_{\mathrm{cr}}-2^{-1}\left(\theta_1+\tilde{\theta}_2-\lambda_0-\lambda_1\right))$ 的渐近方差。与定理1(b)的证明类似，我们有

$$\sqrt{n}\left(\hat{\theta}_{\mathrm{cr}}-\frac{1}{2}\left(\theta_1+\tilde{\theta}_2-\lambda_0-\lambda_1\right)\right)\xrightarrow{d}N\left(0,\frac{\mathrm{Var}(Y_{i1}^{(1)}-Y_{i2}^{(0)})}{4\pi_1}+\frac{\mathrm{Var}(Y_{i2}^{(1)}-Y_{i1}^{(0)})}{4\pi_0}\right).$$

S3.3 生成相关二元随机变量

引理 2。 考虑一个二元变量 $Z_1\sim$ 伯努利分布(p_1)，其中 $\Pr(Z_1=1)=p_1$。给定某个$p_2\in(0,1)$且

$$\begin{aligned}-1\leq\max&\left\{-\sqrt{\frac{p_1p_2}{(1-p_1)(1-p_2)}},-\sqrt{\frac{(1-p_1)(1-p_2)}{p_1p_2}}\right\}<\rho\\&<\min\left\{\sqrt{\frac{p_1(1-p_2)}{p_2(1-p_1)}},\sqrt{\frac{p_2(1-p_1)}{p_1(1-p_2)}}\right\}\leq 1.\end{aligned}$$

如果 $Z_2|Z_1=1\sim$伯努利分布 $\binom{s}{p_1}$ 且 $Z_2|Z_1=0\sim$伯努利分布 $\binom{p_2-s}{1-p_1}$ 其中 $s=\rho\sqrt{p_1(1-p_1)p_2(1-p_2)}+p_1p_2$，那么 $E(Z_2)=p_2$ 且 $\mathrm{Corr}(Z_1,\ Z_2)=\rho$。

证明。首先，很容易证明，当 ρ 满足上述条件时，可推出

$$\begin{cases}-1\le\rho\le 1\\ 0<\frac{s}{p_1}<1\\ 0<\frac{p_2-s}{1-p_1}<1\end{cases}$$

然后，我们也可以证明

$$\begin{aligned}\Pr(Z_2=1|Z_1=1)&=\frac{\Pr(Z_2=1,Z_1=1)}{\Pr(Z_1=1)}=\frac{s}{p_1}\Rightarrow\Pr(Z_2=1,Z_1=1)=s\\ \Pr(Z_2=1|Z_1=0)&=\frac{\Pr(Z_2=1,Z_1=0)}{\Pr(Z_1=0)}=\frac{p_2-s}{1-p_1}\Rightarrow\Pr(Z_2=1,Z_1=0)=p_2-s.\end{aligned}$$

这意味着

$$\Pr(Z_2=1)=\Pr(Z_2=1,Z_1=1)+\Pr(Z_2=1,Z_1=0)=p_2=E(Z_2)$$

于是我们有 $\mathrm{Var}(Z_2)=p_2(1-p_2)$。鉴于 $\mathrm{Var}(Z_1)=p_1(1-p_1)$，我们有

$$\begin{aligned}\mathrm{Corr}(Z_1,Z_2)&=\frac{\mathrm{Cov}(Z_1,Z_2)}{\sqrt{\mathrm{Var}(Z_1)\mathrm{Var}(Z_2)}}=\frac{E(Z_1Z_2)-p_1p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}}\\ &=\frac{\Pr(Z_1=1,Z_2=1)-p_1p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}}=\frac{s-p_1p_2}{\sqrt{p_1(1-p_1)p_2(1-p_2)}}=\rho\end{aligned}$$

□