Carry-Over Effects, Sample Size and Power Consideration in Crossover Trials

Danni Shi and Ting Ye*

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.

November 2022

Abstract

The crossover trial is highly efficient when there is no carry-over effect. A washout period is often designed to reduce the impact of biological carry-over effects. However, the carry-over effect remains an outstanding concern when a washout period is unethical due to the characteristics of the study or cannot sufficiently diminish the persisting impact of period 1. This can occur in implementation trials where the carry-over effect is often behavioral rather than biological.

In this paper, we first give a causal inference reasoning to the crossover design under a potential outcome framework. Under the setting with carry-over effects, our key observation is when the carry-over effect satisfies a sign condition, the basic estimator underestimates the true treatment effect, which does not inflate the type I error of one-sided tests but negatively impacts the statistical power of the trial. This leads to a power trade-off between the crossover and the parallel designs, and we derive the condition under which the crossover design does not bring type I error inflation and is still more powerful than the parallel design. We also propose a covariate adjustment method that improves efficiency, and we present its asymptotic properties. We then use the simulation data to examine the power trade-off and efficiency gain and apply the proposed methods to a real data example.

Keywords: Crossover trials; Causal inference; Covariate adjustment; Implementation science

*tingye1@uw.edu. This work was supported by the HIV Prevention Trials Network (HPTN) and NIH grant: NIAID 5 UM1 AI068617.

1 Introduction

In a crossover trial, all patients could serve as their own control, which eliminates the between-patient variability and economizes sample size, and thus it is often more efficient than a parallel design that uses data of period 1 only. One key assumption for the crossover design is there is no *carry-over effect*: that is, the treatment applied at period 1 does not interfere with the outcome at period 2. So a crossover design is most common for treatments whose effect vanishes when discontinued and for non-absorbing endpoints. Examples include early phase trials (e.g., pharmacokinetic studies and dose finding studies) and phase III studies with chronic conditions (e.g., hypertension, pain, and asthma); see Jones and Lewis (1995) for a review.

A washout period is often inserted between period 1 and 2 to effectively reduce the *biological carry-over effects* engendered by the treatment applied at period 1. However, when a washout is unethical due to the characteristics of the study or cannot sufficiently diminish persisting impact of period 1, the carry-over effect would remain an outstanding concern. An inspirational example is HPTN 104, a recently developed HIV prevention trial protocol: a dual prevention pill (DPP, a daily oral tenofovir disoproxil fumarate–emtricitabine (TDF–FTC) used as pre-exposure prophylaxis (PrEP), plus a combined oral contraceptive) is compared against a two-pill regime using a two-period crossover design. Consistent adherence to TDF-FTC is often challenged by patients' safety concerns (e.g., worries about side effects), fear for stigma, and negative clinic experiences (e.g., being asked highly personal questions), and the hypothesis to be tested here is that the DPP can increase the adherence rate to TDF–FTC. In HIV prevention trials, it is often unethical to insert a washout period and such non-biological carry-over effect is hard to be eliminated even if a washout could be designed. For instance, in HPTN 104, at period 1, if participants taking the DPP are more likely to develop adherence habits, such drug-use habits may carry over into period 2 and affect the adherence for the two-pill regime at period 2. This type of carry-over effect is non-biological and is hard to be eliminated by wash-out periods, and we term it as the *behavioral carry-over effect*. This behavioral carry-over effect can be an outstanding consideration in other HIV prevention trials (e.g., MTN 034 study (NCT03593655), TRIO study (Minnis et al., 2018)) and more broadly in implementation trials (Harichund et al., 2019) that applied crossover designs.

The presence of carry-over effects can bias the estimation for treatment effect. This motivated Grizzle (1965)'s two-stage procedure, which first tests whether the carry-over effect is significant, and the testing result determines whether the period 2 data are used in analysis. Grizzle's method

was criticized in Brown Jr (1980) as it inflates the type I error and decreases the power of the crossover trial. Another type of methods is to model the carry-over effect (Brown Jr, 1980; Laird et al., 1992; Jones and Donev, 1996; Kunert and Stufken, 2002; Bailey and Kunert, 2006) but is sensitive to model misspecification. In this paper, we first give a causal inference reasoning to crossover trials under a potential outcome framework (Neyman, 1923; Rubin, 1974). We then derive asymptotic properties of the basic estimator, and show that the crossover design is typically more efficient than the parallel one when there is no carry-over effect. Under the setting with carry-over effects, our key observation is when the carry-over effect satisfies a sign condition, the basic estimator underestimates the true treatment effect, which does not inflate the type I error of one-sided tests but negatively impacts the power. This leads to a power trade-off between the crossover and the parallel designs, and we derive the condition under which the crossover design does not bring type I error inflation and is still more powerful than the parallel design. We also propose a covariate adjustment method that improves efficiency and present its asymptotic properties. We then use the simulation data to examine the power trade-off and efficiency gain and apply the proposed methods to a real pharmacological data example regarding headache pain relief.

2 Crossover trials with no carry-over effect

2.1 Setup and assumptions

Consider a crossover trial for two treatments in two periods. A sample of n subjects are randomly allocated to two treatment sequences, where $A_i = 1$ denotes that subject i first receives treatment 1 and then treatment 0, and $A_i = 0$ denotes the reverse order. Let $Y_{i1}^{(j)}$ be the potential outcome at time 1 had the subject been exposed to treatment j at time 1, for $j = 0, 1$. Let $Y_{i2}^{(jk)}$ be the potential outcome at time 2 had the subject been exposed to treatment j at time 1 and treatment k at time 2, for $j, k = 0, 1$. The observed outcome for subject i at time t is Y_{it}. Throughout the article, we make the consistency assumption that links the observed outcome to the potential outcomes: for $A_i = 1, Y_{i1} = Y_{i1}^{(1)}$ and $Y_{i2} = Y_{i2}^{(10)}$; for $A_i = 0, Y_{i1} = Y_{i1}^{(0)}$ and $Y_{i2} = Y_{i2}^{(01)}$. Let X_i be the observed baseline covariate for subject i. We assume that $(A_i, X_i, Y_{i1}^{(j)}, Y_{i2}^{(jk)}, j, k = 0, 1), i = 1, \ldots, n$ are independent and identically distributed with finite second order moments.

Simple randomization assigns subjects to the two treatment sequences completely at random. This is summarized in Assumption 1.

Assumption 1 (Randomization). $A_i \perp (X_i, Y_{i1}^{(j)}, Y_{i2}^{(jk)})$ *for* $j, k = 0, 1$. $E(A_i) = \pi_1$ *where* $0 < \pi_1 < 1$ *is known and* $\pi_0 = 1 - \pi_1$.

Assumptions 2 is the key assumption typically imposed in crossover trials. It says that the treatment at time 1 does not have a direct effect on the outcome at time 2; see Figure 1 for an illustration. Many crossover trials would plan a sufficiently long wash-out period between the two time periods to make this assumption more plausible.

Assumption 2 (No carry-over effect). *For* $k = 0, 1$, $Y_{i2}^{(0k)} = Y_{i2}^{(1k)} := Y_{i2}^{(k)}$ *almost surely.*

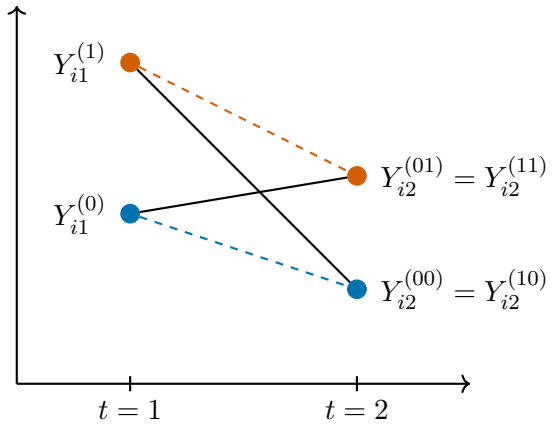

Figure 1: Six potential outcomes with no carry-over effect, i.e., when Assumptions 2 holds.

Under Assumption 2, we can simply let $Y_{i2}^{(k)}, k = 0, 1$ denote the potential outcome at time 2. We are interested in the average treatment effects $\theta_1 = E(Y_{i1}^{(1)} - Y_{i1}^{(0)})$ and $\theta_2 = E(Y_{i2}^{(1)} - Y_{i2}^{(0)})$. Note that it is common to assume that the treatment effect time-invariant (Hills and Armitage, 1979). However, as shown in Theorem 1 below, a time-invariant treatment effect assumption is not necessary for crossover trials.

2.2 The basic estimator

Using the full crossover period, we can calculate the difference in outcome for every subject under treatment 1 and treatment 0. The *basic estimator* is commonly used, which first calculates the arm-specific average outcome difference, and then averages the two means from the two arms (Hills and Armitage, 1979; Senn, 1994; Kim et al., 2021, Chapter 8):

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

where $\bar{\Delta}_a = n_a^{-1} \sum_{A_i=a} \Delta_i$ is the average of Δ_i's for subjects with $A_i = a$, $\Delta_i = Y_{i1} - Y_{i2}$, and n_a is the number of subjects with $A_i = a$.

Theorem 1 summarizes the statistical properties of the basic estimator $\hat{\theta}_{\text{cr}}$. This and all other proofs are in the supplementary materials.

Theorem 1. *Under Assumptions 1-2,*
(a) $E(\hat{\theta}_{\text{cr}}) = 2^{-1}(\theta_1 + \theta_2)$, *where* $\theta_t = E(Y_{it}^{(1)} - Y_{it}^{(0)})$.
(b) $\sqrt{n}\{\hat{\theta}_{\text{cr}} - 2^{-1}(\theta_1 + \theta_2)\} \xrightarrow{d} N\left(0, \sigma_{\text{cr}}^2\right)$, *where* $\sigma_{\text{cr}}^2 = (4\pi_1)^{-1}\text{Var}(Y_{i1}^{(1)} - Y_{i2}^{(0)}) + (4\pi_0)^{-1}\text{Var}(Y_{i1}^{(0)} - Y_{i2}^{(1)})$.

For Theorem 1(a), we show that the expected change in outcome for $A_i = 1$ is the average treatment effect at time 1 minus the expected change in outcome in the absence of the treatment, i.e., $E(Y_{i1} - Y_{i2} \mid A_i = 1) = E(Y_{i1}^{(1)} - Y_{i2}^{(0)}) = E(Y_{i1}^{(1)} - Y_{i1}^{(0)} + Y_{i1}^{(0)} - Y_{i2}^{(0)}) = \theta_1 - \tau$, where $\tau = E(Y_{i2}^{(0)} - Y_{i1}^{(0)})$ denotes the expected change in outcome in the absence of the treatment. Similarly, the expected change in outcome for $A_i = 0$ is the average treatment effect plus the expected change in outcome in the absence of the treatment, i.e., $E(Y_{i2} - Y_{i1} \mid A_i = 0) = E(Y_{i2}^{(1)} - Y_{i1}^{(0)}) = E(Y_{i2}^{(1)} - Y_{i2}^{(0)} + Y_{i2}^{(0)} - Y_{i1}^{(0)}) = \theta_2 + \tau$. At first sight, it appears that both the group-specific average change in outcome are biased by τ, the time trend not due to the treatment. However, randomization balances out the effect of time trend not due to the treatment, and thus the overall average change in outcome remains unbiased for $(\theta_1 + \theta_2)/2$. Moreover, with a fixed sample size n, Theorem 1(b) guides the optimal choice of π_1 to minimize the variance. For example, if $\text{Var}(Y_{i1}^{(1)} - Y_{i2}^{(0)}) = \text{Var}(Y_{i1}^{(0)} - Y_{i2}^{(1)})$, then the optimal choice is equal allocation with $\pi_1 = 1/2$.

For statistical inference, σ_{cr}^2 can be estimated via

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

where $S_{\Delta 1}^2$ and $S_{\Delta 0}^2$ are respectively the sample variance of $\Delta_i = Y_{i1} - Y_{i2}$ for subjects under $A_i = 1$ and $A_i = 0$. This variance estimator is consistent as $n \to \infty$.

There is another estimator that looks similar to the basic estimator $\hat{\theta}_{\text{cr}}$: $\hat{\theta}_{\text{cr}}^* = n^{-1} \sum_{i=1}^n \{A_i(Y_{i1} - Y_{i2}) + (1 - A_i)(Y_{i2} - Y_{i1})\}$. These two estimators $\hat{\theta}_{\text{cr}}$ and $\hat{\theta}_{\text{cr}}^*$ are the same under a randomization scheme that enforces $n_1 = n_0$, but they are not the same in other cases (e.g., under the simple randomization considered in this article). Specifically, under simple randomization with equal allocation ($\pi_1 = 1/2$), $\hat{\theta}_{\text{cr}}^*$ is also unbiased for $(\theta_1 + \theta_2)/2$; however, the asymptotic variance of $\sqrt{n}\{\hat{\theta}_{\text{cr}}^* - 2^{-1}(\theta_1 + \theta_2)\}$ equals $2^{-1}\text{Var}(Y_{i1}^{(1)} - Y_{i2}^{(0)}) + 2^{-1}\text{Var}(Y_{i1}^{(0)} - Y_{i2}^{(1)}) + 4^{-1}(\theta_1 - \theta_2 - 2\tau)^2$,

which is larger than the variance of $\hat{\theta}_{\text{cr}}$. This additional variance component is due to the random effect of time trend and the treatment effect heterogeneity at the two time points. Under simple randomization with unequal allocation, $\hat{\theta}^*_{\text{cr}}$ is biased due to the time effect τ. This important point is also discussed in Hills and Armitage (1979). Therefore, we do not consider this estimator in the rest of the article.

2.3 Efficiency comparison between a crossover and parallel-group design

When there is no carry-over effect, i.e., when Assumption 2 holds, it is well known that a crossover design is typically more efficient than a parallel-group design. This is because each subject can serve as their own control, which cuts down half the sample size, and the within-subject comparison can further remove the inter-subject variability (Jones and Kenward, 1989). In this section, we provide a formal efficiency comparison between a crossover and parallel-group design under Assumption 2.

Suppose a randomized controlled trial is being planned to demonstrate the superiority or non-inferiority of an investigational treatment. First, consider the crossover design, the null hypothesis $H_0 : 2^{-1}(\theta_1+\theta_2) = \theta^*$ versus the alternative hypothesis $H_A : 2^{-1}(\theta_1+\theta_2) > \theta^*$ for some pre-specified θ^*. (Under the test for superiority, $\theta^* = 0$; for non-inferiority, $\theta^* > 0$.) The test statistic based on the basic estimator is $T_{\text{cr}} = \sqrt{n}(\hat{\theta}_{\text{cr}} - \theta^*)/\hat{\sigma}_{\text{cr}}$. From Theorem 1, $T_{\text{cr}} \xrightarrow{d} N(0,1)$ under H_0, and thus, we reject H_0 if and only if $T_{\text{cr}} > z_{1-\alpha}$, where α is the significance level and $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standard normal distribution. Under the local alternative $\sqrt{n}\{2^{-1}(\theta_1+\theta_2) - \theta^*\} \to \gamma_{\text{cr}}$ with a constant $\gamma_{\text{cr}} > 0$, the power of T_{cr} is

$$\text{Power}_{\text{cr}} \approx \Phi\left(-z_{1-\alpha} + \frac{\gamma_{\text{cr}}}{\sigma_{\text{cr}}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\approx$ denotes asymptotic approximation.

Now suppose we only use the data at time 1, which is effectively a parallel-group design. The parallel-group counterpart to $\hat{\theta}_{\text{cr}}$ is a simple mean difference $\hat{\theta}_{\text{pr}} = \frac{1}{n_1}\sum_{i=1}^{n} A_i Y_{i1} - \frac{1}{n_0}\sum_{i=1}^{n}(1 - A_i)Y_{i1}$, where n_a is the number of subjects under $A_i = a$. Under Assumption 1, we have $E(\hat{\theta}_{\text{pr}}) = \theta_1$ and $\sqrt{n}(\hat{\theta}_{\text{pr}} - \theta_1) \xrightarrow{d} N(0, \sigma^2_{\text{pr}})$, where $\sigma^2_{\text{pr}} = 2\text{Var}(Y^{(1)}_{i1}) + 2\text{Var}(Y^{(0)}_{i1})$. For testing the null hypothesis $H_0 : \theta_1 = \theta^*$ versus the alternative hypothesis $H_A : \theta_1 > \theta^*$, the test statistic is $T_{\text{pr}} = \sqrt{n}(\hat{\theta}_{\text{pr}} - \theta^*)/\hat{\sigma}_{\text{pr}}$, where $\hat{\sigma}^2_{\text{pr}}$ is a consistent estimator for σ^2_{pr}. Under the local alternative

$\sqrt{n}\{\theta_1 - \theta^*\} \to \gamma_{\text{pr}}$ with a constant $\gamma_{\text{pr}} > 0$,

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

Let $\text{Power}_{\text{cr}} = \text{Power}_{\text{cr}} = 1 - \beta$, we obtain the required sample sizes to achieve $1 - \beta$ power using the two designs:

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

Suppose $\theta_1 = \theta_2$, then the two above tests for the two designs are inherently the same, and the ratio of the sample sizes to achieve the same power between two tests is also the Pitman asymptotic relative efficiency and can be expressed as:

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

For illustration, consider a simple case where $\theta_1 = \theta_2$, $\text{Var}(Y_{it}^{(j)}) = \sigma^2$ for $t = 1, 2$ and $j = 0, 1$, and $\text{Cov}(Y_{i1}^{(j)}, Y_{i2}^{(1-j)}) = \rho\sigma^2$, where $\rho \geq 0$ is the intraclass correlation coefficient (ICC). Then $\sigma_{\text{cr}}^2 = 2(1 - \rho)\sigma^2$ and $\sigma_{\text{pr}}^2 = 4\sigma^2$, and $n_{\text{cr}}/n_{\text{pr}} = (1 - \rho)/2$. This means that the crossover design only requires $(1 - \rho)/2$ of the sample size required by the parallel-group design, which is the key advantage of the crossover design over the parallel-group design.

3 Crossover trials with carry-over effect

3.1 Setup and assumptions

When there exist carry-over effects in a crossover trial, the treatment at time 1 may interfere with the outcome at time 2; see Figure 2 for an illustration of the six potential outcomes with carry-over effect. In this case, Assumption 1 still holds by the act of randomization, while Assumption 2 is violated.

We present a way of parameterizing the six potential outcome means in Table 1 that directly extends the parameterization used in Section 2. In Table 1, $\tilde{\theta}_2 = E(Y_{i2}^{(11)} - Y_{i2}^{(00)})$ denotes the average treatment effect at time 2, which compares the potential outcome had one stays on treatment 1 to the potential outcome had one stays on treatment 0, $\tilde{\tau} = E(Y_{i2}^{(00)} - Y_{i1}^{(0)})$ denotes the expected change in outcome had one stays on treatment 0, $\lambda_0 = E(Y_{i2}^{(10)} - Y_{i2}^{(00)})$ and $\lambda_1 = E(Y_{i2}^{(11)} - Y_{i2}^{(01)})$

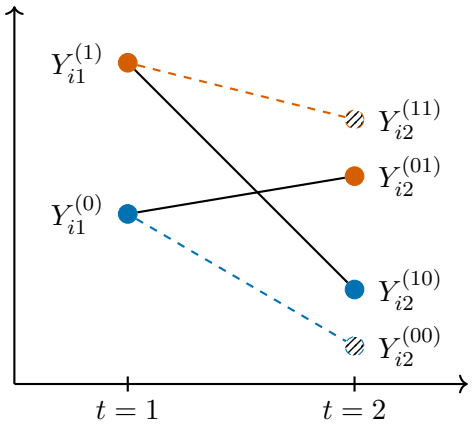

Figure 2: Six potential outcomes with carry-over effect, i.e., when Assumptions 2 is violated.

are the expected carry-over effect under the two treatment regimes.

Table 1: Parameterization of the six potential outcome means with and without Assumption 2. The two shadowed rows correspond to the potential outcomes that are never observed.

Potential outcome means	With Assumption 2	Without Assumption 2
$E(Y_{i1}^{(0)})$	μ	μ
$E(Y_{i1}^{(1)})$	$\mu + \theta_1$	$\mu + \theta_1$
$E(Y_{i2}^{(00)})$	$\mu + \tau$	$\mu + \tilde{\tau}$
$E(Y_{i2}^{(10)})$	$\mu + \tau$	$\mu + \tilde{\tau} + \lambda_0$
$E(Y_{i2}^{(11)})$	$\mu + \tau + \theta_2$	$\mu + \tilde{\tau} + \tilde{\theta}_2$
$E(Y_{i2}^{(01)})$	$\mu + \tau + \theta_2$	$\mu + \tilde{\tau} + \tilde{\theta}_2 - \lambda_1$

3.2 The basic estimator

Theorem 2 derives the statistical properties of the basic estimator $\hat{\theta}_{\text{cr}}$ defined in (1) without Assumption 2.

Theorem 2. *Under Assumption 1,*

(a) $E(\hat{\theta}_{\text{cr}}) = 2^{-1}(\theta_1 + \tilde{\theta}_2 - \lambda_0 - \lambda_1)$.

(b) $\sqrt{n}\{\hat{\theta}_{\text{cr}} - 2^{-1}(\theta_1 + \tilde{\theta}_2 - \lambda_0 - \lambda_1)\} \xrightarrow{d} N\left(0, \tilde{\sigma}_{\text{cr}}^2\right)$, *where* $\tilde{\sigma}_{\text{cr}}^2 = (4\pi_1)^{-1}\text{Var}(Y_{i1}^{(1)} - Y_{i2}^{(10)}) + (4\pi_0)^{-1}\text{Var}(Y_{i2}^{(01)} - Y_{i1}^{(0)})$.

Similar to the proof of Theorem 1(a), we show that the expected change in outcome for $A_i = 1$ is $E(Y_{i1} - Y_{i2} \mid A_i = 1) = E(Y_{i1}^{(1)} - Y_{i2}^{(10)}) = E(Y_{i1}^{(1)} - Y_{i1}^{(0)} + Y_{i1}^{(0)} - Y_{i2}^{(00)} + Y_{i2}^{(00)} - Y_{i2}^{(10)}) = \theta_1 - \tilde{\tau} - \lambda_0$, that for $A_i = 0$ is $E(Y_{i2} - Y_{i1} \mid A_i = 0) = E(Y_{i2}^{(01)} - Y_{i1}^{(0)}) = E(Y_{i2}^{(01)} - Y_{i2}^{(11)} + Y_{i2}^{(11)} - Y_{i2}^{(00)} + Y_{i2}^{(00)} - Y_{i1}^{(0)}) = \tilde{\theta}_2 + \tilde{\tau} - \lambda_1$. Hence, the overall average change in outcome is $2^{-1}(\theta_1 + \tilde{\theta}_2 - \lambda_0 - \lambda_1)$.

Hence, Theorem 2(a) implies that $\hat{\theta}_{\rm cr}$ is still an estimator of a treatment effect $2^{-1}(\theta_1+\tilde{\theta}_2)$ under a population-level no carry-over effect assumption ($E(Y_{i2}^{(1k)}) = E(Y_{i2}^{(0k)})$ for $k = 0, 1$), which is weaker than the individual-level no carry-over effect assumption stated in Assumption 2. It is also straightforward to verify that Theorem 1 is a special case of Theorem 2 under Assumption 2.

Note that $\hat{\sigma}^2_{\rm cr}$ defined in (2) is still a consistent estimator of $\tilde{\sigma}^2_{\rm cr}$ without Assumption 2.

3.3 Type I error and power analysis

With possible carry-over effects, the first question one would ask is whether using $T_{\rm cr}$ to analyze crossover trials leads to type I error inflation. To answer this question, consider the null hypothesis $H_0 : 2^{-1}(\theta_1+\tilde{\theta}_2) = \theta^*$ versus the alternative hypothesis $H_A : 2^{-1}(\theta_1+\tilde{\theta}_2) > \theta^*$ for some pre-specified θ^*. The type I error rate of $T_{\rm cr} = \sqrt{n}(\hat{\theta}_{\rm cr} - \theta^*)/\hat{\sigma}_{\rm cr}$ is

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

When $\lambda_0+\lambda_1 = 0$, $\hat{\theta}_{\rm cr}$ is an unbiased estimator for the treatment effect $2^{-1}(\theta_1+\tilde{\theta}_2)$ and the type I error rate of $T_{\rm cr}$ is α. When $\lambda_0+\lambda_1 > 0$, $\hat{\theta}_{\rm cr}$ under-estimates the treatment effect $2^{-1}(\theta_1+\tilde{\theta}_2)$ and the type I error rate is less than α, meaning that the test is conservative. When $\lambda_0+\lambda_1 < 0$, $\hat{\theta}_{\rm cr}$ over-estimates the treatment effect $2^{-1}(\theta_1+\tilde{\theta}_2)$ and the type I error rate is larger than α, meaning that the test is invalid. Therefore, even with carry-over effects, the crossover design can still control the type I error rate of a one-sided test when $\lambda_0+\lambda_1 \geq 0$.

However, carry-over effects that do not inflate type I error rate can have a negative impact on the power. Specifically, with possible carry-over effects and under the local alternative $\sqrt{n}\{2^{-1}(\theta_1+\tilde{\theta}_2)-\theta^*\} \to \gamma_{\rm cr}$ with a constant $\gamma_{\rm cr} > 0$, the power of $T_{\rm cr}$ is

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

The required sample size to achieve $1-\beta$ power is

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

For illustration, suppose that $\theta_1 = \tilde{\theta}_2 = \theta_{\text{Alt}} > 0$ and $\theta^* = 0$ (a test of superiority), the Pitman asymptotic relative efficiency between the crossover and parallel-group design is

$$\frac{\tilde{n}_{\text{cr}}}{n_{\text{pr}}} = \frac{\tilde{\sigma}^2_{\text{cr}}}{\sigma^2_{\text{pr}}} \cdot \left\{1 - \frac{\lambda_0 + \lambda_1}{2\ \theta_{\text{Alt}}}\right\}^{-2} \tag{5}$$

When $\theta_1 + \tilde{\theta}_2 > \lambda_0 + \lambda_1 \geq 0$, i.e., the carry-over effect is non-negative to control the type I error rate and is also smaller than the treatment effect, the difference between n_{pr} and $\tilde{n}_{\text{cr}}$ equals

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

where c is a positive constant. Therefore, in order for the crossover design to reduce sample sizes, we need the carry-over effect to be small. Specifically, similar to our discussion at the end of Section 2: when $\tilde{\sigma}_{\text{cr}}/\sigma_{\text{pr}} = \sqrt{(1-\rho)/2}$, we need $\frac{2^{-1}(\lambda_0+\lambda_1)}{\theta_{\text{Alt}}} < 1 - \sqrt{(1-\rho)/2}$. When $\rho = 0.3, 0.5, 0.7$, we need to have $\frac{1}{2}(\lambda_0 + \lambda_1) > 0$ and less than $0.41\theta_{\text{Alt}}, 0.5\theta_{\text{Alt}}, 0.61\theta_{\text{Alt}}$ respectively so that the type I error is not inflated and the crossover design is more powerful than the parallel one. This small carry-over effect condition may be plausible in many scenarios because carry-over effects are usually relatively small compared to the treatment effect of primary interest. Therefore, crossover design may be more powerful than a parallel-group design in many cases even when carry-over effects exist. In Section B of the appendix, we discuss how to control the type I error when $\lambda_0 + \lambda_1 < 0$ using a sensitivity analysis approach (Rosenbaum, 2020).

4 Covariate adjustment for crossover trials

Covariate adjustment is a statistical method with high potential to improve precision for many clinical trials (FDA, 2021). It often uses a *working* model between outcomes and covariates, but its estimand is the same as under unadjusted methods and its inference does not rely on the working model being correctly specified. Covariate adjustment for parallel trials has been extensively studied; particularly, it has been established that using a linear model as a working model for covariate adjustment leads to *guaranteed efficiency gain* regardless of the model being mispecified or not (Yang and Tsiatis, 2001; Lin, 2013; Ye et al., 2022a,b). These recent results have not been extended to crossover trials, although covariate adjustment is broadly recommended for crossover trials (Metcalfe, 2010; Mehrotra, 2014; Jemielita et al., 2016).

To use the baseline covariate $\boldsymbol{X}_i$, we propose the following covariate-adjusted estimator

$$\hat{\theta}_{\text{cr,adj}} = \frac{\{\bar{\Delta}_1 - \hat{\boldsymbol{\beta}}_1^T(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}})\} - \{\bar{\Delta}_0 - \hat{\boldsymbol{\beta}}_0^T(\bar{\boldsymbol{X}}_0 - \bar{\boldsymbol{X}})\}}{2}, \tag{6}$$

where $\bar{\boldsymbol{X}}$ is the sample mean of all $\boldsymbol{X}_i$'s, $\bar{\boldsymbol{X}}_a$ is the sample mean of $\boldsymbol{X}_i$'s from subjects with $A_i = a$, and

$$\hat{\boldsymbol{\beta}}_a = \left\{ \sum_{i:A_i=a} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_a)(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_a)^T \right\}^{-1} \sum_{i:A_i=a} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_a)\Delta_i$$

is the least squares estimator of $\boldsymbol{\beta}_a$ in fitting the linear working model $E(\Delta_i \mid A_i = a, \boldsymbol{X}_i) = \mu_a + \boldsymbol{\beta}_a^T \boldsymbol{X}_i$ using subjects with $A_i = a$.

The following heuristics reveal why covariate adjustment does not change the estimand, often gains but never hurts efficiency even when the the linear working model is wrong. As randomization balances the covariate distribution, both $\bar{\boldsymbol{X}}_a$ and $\bar{\boldsymbol{X}}$ estimate the same quantity and thus, $\hat{\boldsymbol{\beta}}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$ is an "estimator" of zero. Hence, $\bar{\Delta}_a - \hat{\boldsymbol{\beta}}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$ and $\bar{\Delta}_a$ correspond to the same estimand. In addition, as $n \to \infty$, $\hat{\boldsymbol{\beta}}_a$ converges to $\boldsymbol{\beta}_a = \text{Var}(\boldsymbol{X}_i)^{-1}\text{Cov}(\boldsymbol{X}_i, \Delta_i \mid A_i = a)$ in probability, regardless of the linear working model is correct or not. Hence, $\bar{\Delta}_a - \hat{\boldsymbol{\beta}}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$ is asymptotically equivalent to $\bar{\Delta}_a - \boldsymbol{\beta}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$. Note that

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

Consequently, $\bar{\Delta}_a - \hat{\boldsymbol{\beta}}_a^T(\bar{\boldsymbol{X}}_a - \bar{\boldsymbol{X}})$ has a smaller asymptotic variance than $\bar{\Delta}_a$. These results are formally stated in Theorem 3; see Appendix A.2 for proof.

Theorem 3. *Under Assumptions 1,*
(a) $\sqrt{n}\{\hat{\theta}_{\text{cr,adj}} - 2^{-1}(\theta_1 + \tilde{\theta}_2 - \lambda_0 - \lambda_1)\} \xrightarrow{d} N\left(0, \tilde{\sigma}^2_{\text{cr,adj}}\right)$, *where* $\tilde{\sigma}^2_{\text{cr,adj}} = (4\pi_1)^{-1}\text{Var}(Y_{i1}^{(1)} - Y_{i2}^{(10)} - \boldsymbol{\beta}_1^T\boldsymbol{X}_i) + (4\pi_0)^{-1}\text{Var}(Y_{i1}^{(0)} - Y_{i2}^{(01)} - \boldsymbol{\beta}_0^T\boldsymbol{X}_i) + 4^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T\text{Var}(\boldsymbol{X})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$, *and* $\boldsymbol{\beta}_a = \text{Var}(\boldsymbol{X}_i)^{-1}\text{Cov}(\boldsymbol{X}_i, \Delta_i \mid A_i = a)$.
(b) Moreover, $\tilde{\sigma}^2_{\text{cr,adj}} - \tilde{\sigma}^2_{\text{cr}} = -(4\pi_1\pi_0)^{-1}(\pi_0\boldsymbol{\beta}_1 + \pi_1\boldsymbol{\beta}_0)^T\text{Var}(\boldsymbol{X})(\pi_0\boldsymbol{\beta}_1 + \pi_1\boldsymbol{\beta}_0) \leq 0$.

From Theorem 3(b), the asymptotic variance of $\hat{\theta}_{\text{cr,adj}}$ is no larger than that of $\hat{\theta}_{\text{cr}}$. In fact, Theorem 1 of Ye et al. (2022a) implies a stronger result that $\hat{\theta}_{\text{cr,adj}}$ has the smallest asymptotic variance among all linearly adjusted estimators of the form $2^{-1}\big[\{\bar{\Delta}_1 - \boldsymbol{b}_1^T(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}})\} - \{\bar{\Delta}_0 -$

$\boldsymbol{b}_0^T(\bar{\boldsymbol{X}}_0 - \bar{\boldsymbol{X}})\}\big]$, where $\boldsymbol{b}_0, \boldsymbol{b}_1$ are any fixed or random vectors that have the same dimension as $\boldsymbol{X}_i$.

A consistent estimator for $\tilde{\sigma}^2_{\text{cr,adj}}$ is

$$\hat{\sigma}^2_{\text{cr,adj}} = \frac{1}{4\pi_1} S^2_{\Delta 1,\text{adj}} + \frac{1}{4\pi_0} S^2_{\Delta 0,\text{adj}} + \frac{1}{4}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)^T \hat{\Sigma}_X (\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0), \tag{7}$$

where $S^2_{\Delta a,\text{adj}}$ is the sample variance of $Y_{i1} - Y_{i2} - \hat{\boldsymbol{\beta}}_a^T \boldsymbol{X}_i$ based on data under $A_i = a$, for $a = 0, 1$, $\hat{\Sigma}_X$ is the sample covariance matrix of $\boldsymbol{X}_i$ based on all samples. One can easily construct a Z-test based on the covariate-adjusted estimator $\hat{\theta}_{\text{cr,adj}}$, which is guaranteed to be more powerful than the unadjusted counterpart T_{cr} from Theorem 3.

5 Power calculations and simulation

To compare the powers under two designs, we consider the following data-generating process:

$$\begin{aligned}
Y_{i1}^{(0)} &= X_{i1} + X_{i2} + X_{i3} + \epsilon_{i1} \\
Y_{i1}^{(1)} &= \theta_1 + X_{i1} + X_{i2} + X_{i3} + \epsilon_{i2} \\
Y_{i2}^{(10)} &= \tilde{\tau} + \lambda_0 + X_{i1} + X_{i2} + bX_{i3} + \epsilon_{i3} \\
Y_{i2}^{(01)} &= \tilde{\tau} + \tilde{\theta}_2 - \lambda_1 + X_{i1} + X_{i2} + bX_{i3} + \epsilon_{i4}
\end{aligned} \tag{8}$$

where $X_{ij}, \epsilon_{ik} \sim N(0,1)$ for $j = 1,2,3$ and $k = 1,2,3,4$. The treatment arm indicator $A_i \sim$ Bernoulli($p = \pi_1 = 1/2$). The causal consistency assumption still holds. The observed data are $(\boldsymbol{X}_i, A_i, Y_{i1}, Y_{i2}), i = 1, \ldots, n$. So $\text{Var}(Y_{i1}^{(0)}) = \text{Var}(Y_{i1}^{(1)}) = 4$, $\text{Var}(Y_{i2}^{(10)}) = \text{Var}(Y_{i2}^{(01)}) = b^2 + 3$, $\text{Cov}(Y_{i1}^{(1)}, Y_{i2}^{(10)}) = \text{Cov}(Y_{i1}^{(0)}, Y_{i2}^{(01)}) = b + 2$, and thus $\sigma^2_{\text{pr}} = 16$ and $\tilde{\sigma}^2_{\text{cr}} = (1-b)^2 + 2$. For the covariate-adjusted estimator, we adjust for $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})^T$. Here, $\text{Var}(\boldsymbol{X}_i)$ is a 3×3 identity matrix and $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = (0, 0, 1-b)^T$. Thus, $\tilde{\sigma}^2_{\text{cr,adj}} = 2$. We set $n = 500, \tilde{\tau} = 0, \alpha = 0.025$, $b = 0$, $\theta^* = 0$.

Figure 3 shows the type I error rate and power for three one-sided tests $T_{\text{pr}}, T_{\text{cr}}, T_{\text{cr,adj}}$ when $\theta_1 = \tilde{\theta}_2 = \theta \in [0, 0.5]$ and $\lambda_0 = \lambda_1 = \lambda \in \{-0.1, 0, 0.1, 0.3\}$ based on the formula in (3), (4) and its covariate-adjusted counterpart. In this setting, the value of b affects the power of T_{cr} but not the other two tests, so we also consider $b = \frac{1}{3}$ for T_{cr} only. Using 10,000 repetitions for simulation, we find good agreement between formula powers and simulation powers.

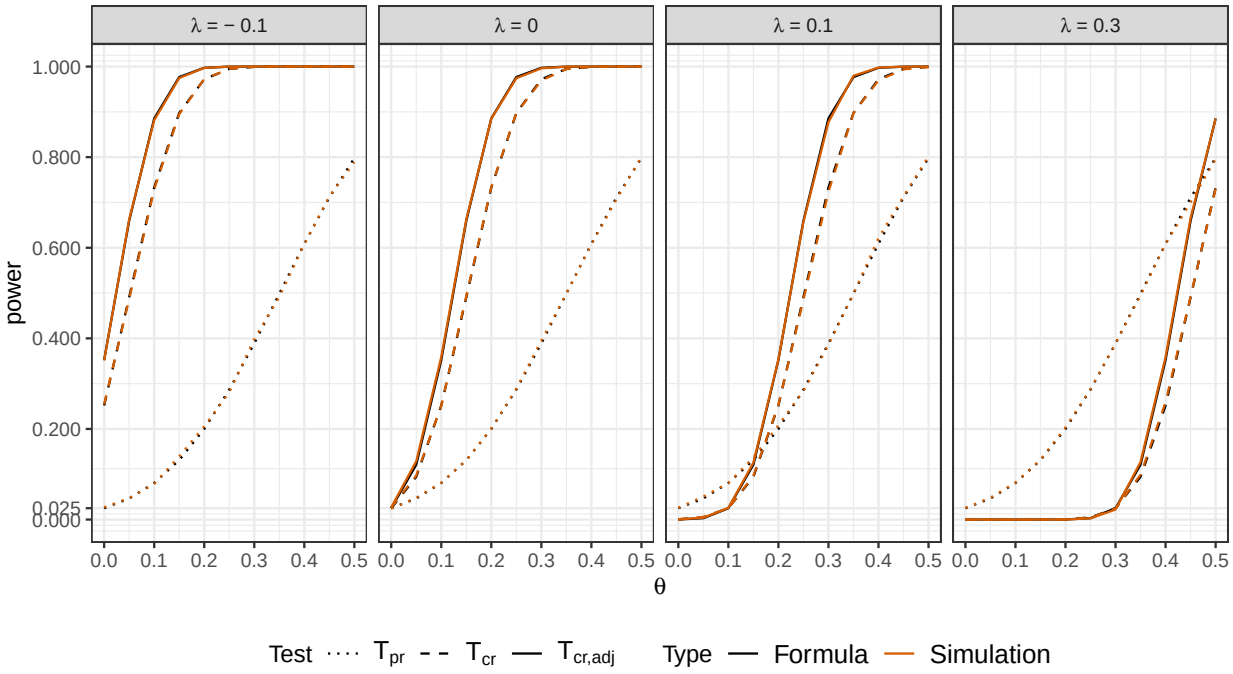

Figure 3: Power curves for three tests calculated by formula and simulations (N=10,000 repetitions) when $\lambda = -0.1, 0, 0.1, 0.3$ and $b = 0, \frac{1}{3}$. Note that the power of T_{pr} and $T_{\text{cr,adj}}$ are unaffected by b.

6 Data Analysis Example

Laird et al. (1992) discussed the analysis of a simple pharmacological crossover trial comparing two active analgesic drugs and placebo for relief of tension headaches. The two active drugs (A, B) have identical active ingredients of 1000 mg of acetaminophen and A contained caffeine: caffeine alone is considered ineffective as an analgesic and here researchers were interested in whether it could work as an adjuvant medication. Placebo arm (P) was included for regulatory purposes. Patients were randomly allocated to 6 treatment sequences (AB, BA, BP, PB, AP, PA) and participated the study in 14 different clinical centers. In each period, two types of headache attacks were treated, and the outcomes were each patient's two scores for their level of pain relief for the two treated headaches. The dataset we use is obtained from Fitzmaurice et al. (2011, Chapter 21.4) and the outcome is the mean of the two measures for pain relief.

Since placebo was applied at period 1, the treatment outcomes at period 2 in PB and PA sequences were unlikely to be affected by carry-over effects. Given that the 95% confidence intervals for the average treatment outcomes at period 2 of the AB $(8.55, 9.75)$ and PB $(8.65, 11.18)$ sequences, BA $(10.18, 11.40)$ and PA $(9.90, 11.81)$ sequences are similar, we assume that treatment A and B do not contain carry-over effects and thus when comparing the treatment effects of A and

Table 2: Descriptive statistics by treatment sequence (mean and standard deviation of the pain relief scores and treatment effects, and intraclass correlation (ρ) between pain relief scores at two periods) for total pain relief scores of headache in periods 1 and 2

		Period 1		Period 2		treatment effect		
Sequence	N	Mean	SD	Mean	SD	Mean	SD	ρ
AB	126	10.196	3.347	9.153	3.429	1.044	4.292	0.198
BA	127	9.581	3.881	10.791	3.530	1.211	4.399	0.298
BP	42	10.333	3.306	8.357	3.944	1.976	2.787	0.718
PB	42	7.464	4.265	9.911	4.183	2.446	3.448	0.667
AP	43	10.477	3.546	7.273	4.451	3.203	4.760	0.308
PA	43	8.366	3.777	10.855	3.204	2.488	3.631	0.469

B, we consider data from AB and BA sequences only.

We adjust for the clinical centers and compare the covariate-adjusted and basic crossover estimators and the parallel estimator (from period 1 data only) for the treatment effect of A from the AB and BA sequences. In Table 3, the two crossover estimators are both significant while parallel estimator is not. The covariate-adjusted crossover estimator has smaller variance than the unadjusted estimator, and the parallel estimator has the largest variance.

Table 3: Adjusted and unadjusted crossover estimators and parallel estimator for the average treatment effect of A versus B, with standard deviations and 95% confidence intervals

Type	Mean	SD	95% CI
Crossover, adjusted	1.139	0.265	(0.621, 1.658)
Crossover, basic	1.127	0.273	(0.592, 1.663)
Parallel	0.616	0.456	(-0.277, 1.509)

7 Conclusion

The crossover trial is often preferred due to its high efficiency as participants could serve as their own controls. In implementation trials, the carry-over effect, especially the behavioral one, may not be able to be reduced due to characteristics of the studies and thus it is an outstanding concern that can bias the estimation for treatment effect and may affect statistical precision. Using a potential outcomes framework, we investigate the impact of carry-over effect in the two-treatment two-period crossover trial, and find that when the carry-over effect $\lambda_1 + \lambda_0$ is non-negative, the basic estimator underestimates the treatment effect, which does not inflate the type I error of one-sided superiority or non-inferiority tests but negatively impacts the power. Furthermore, when

$\lambda_1 + \lambda_0$ is relatively small comparing to the treatment effect, the crossover design can still be more powerful than the parallel design, which uses data from period 1 only. We further apply covariate adjustment in crossover trials for guaranteed efficiency gain. The power trade-off and efficiency gain are examined in a data generating and simulation example and the proposed methods are applied to a real pharmacological data example. All the methods in this article can be implemented using the R package RobinCar, which is available at https://github.com/tye27/RobinCar.

References

Bailey, R. and Kunert, J. (2006). On optimal crossover designs when carryover effects are proportional to direct effects. *Biometrika*, 93(3):613–625.

Brown Jr, B. W. (1980). The crossover experiment for clinical trials. *Biometrics*, pages 69–79.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959). Smoking and lung cancer. *Journal of the National Cancer Institute*, 22:173–203.

FDA (2021). Adjusting for covariates in randomized clinical trials for drugs and biological products. Draft Guidance for Industry. Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, Food and Drug Administration (FDA), U.S. Department of Health and Human Services. May 2021.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis*. John Wiley & Sons Ltd.

Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics*, pages 467–480.

Harichund, C., Karim, Q. A., Kunene, P., Simelane, S., and Moshabela, M. (2019). Hiv self-testing as part of a differentiated hiv testing approach: exploring urban and rural adult experiences from kwazulu-natal, south africa using a cross-over study design. *BMC public health*, 19(1):1–7.

Hills, M. and Armitage, P. (1979). The two-period cross-over clinical trial. *British journal of clinical pharmacology*, 8(1):7.

Jemielita, T., Putt, M., and Mehrotra, D. (2016). Improved power in crossover designs through linear combinations of baselines. *Statistics in Medicine*, 35(30):5625–5641.

Jones, B. and Donev, A. (1996). Modelling and design of cross-over trials. *Statistics in Medicine*, 15(13):1435–1446.

Jones, B. and Kenward, M. G. (1989). *Design and analysis of cross-over trials*. Chapman and Hall/CRC.

Jones, B. and Lewis, J. (1995). The case for cross-over trials in phase iii. *Statistics in Medicine*, 14(9):1025–1038.

Kim, K., Bretz, F., Cheung, Y. K. K., and Hampson, L. V. (2021). *Handbook of statistical methods for randomized controlled trials*. CRC Press.

Kunert, J. and Stufken, J. (2002). Optimal crossover designs in a model with self and mixed carryover effects. *Journal of the American Statistical Association*, 97(459):898–906.

Laird, N. M., Skinner, J., and Kenward, M. (1992). An analysis of two-period crossover designs with carry-over effects. *Statistics in Medicine*, 11(14-15):1967–1979.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Annals of Applied Statistics*, 7(1):295–318.

Mehrotra, D. V. (2014). A recommended analysis for 2× 2 crossover trials with baseline measurements. *Pharmaceutical statistics*, 13(6):376–387.

Metcalfe, C. (2010). The analysis of cross-over trials with baseline measurements. *Statistics in medicine*, 29(30):3211–3218.

Minnis, A. M., Roberts, S. T., Agot, K., Weinrib, R., Ahmed, K., Manenzhe, K., Owino, F., and van der Straten, A. (2018). Young women's ratings of three placebo multipurpose prevention technologies for hiv and pregnancy prevention in a randomized, cross-over study in kenya and south africa. *AIDS and Behavior*, 22(8):2662–2673.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).

Rosenbaum, P. R. (2020). *Design of observational studies (2nd ed.)*. Springer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 6(5):688–701.

Senn, S. (1994). The AB/BA crossover: past, present and future? *Statistical methods in medical research*, 3(4):303–324.

Yang, L. and Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321.

Ye, T., Shao, J., Yi, Y., and Zhao, Q. (2022a). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association*, pages 1–13.

Ye, T., Yi, Y., and Shao, J. (2022b). Inference on the average treatment effect under minimization and other covariate-adaptive randomization methods. *Biometrika*, 109(1):33–47.

Yi, Y., Zhang, Y., Du, Y., and Ye, T. (2022). Testing for treatment effect twice using internal and external controls in clinical trials. *arXiv preprint arXiv:2203.04194*.

A Technical Proofs

A.1 Proof of Theorem 1

(a) We calculate the expectation as follows:

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

(b) Asymptotic normality is straightforward from central limit theorem. In what follows, we derive the asymptotic variance of $\sqrt{n}(\hat{\theta}_{\mathrm{cr}}-2^{-1}(\theta_1+\tilde{\theta}_2-\lambda_0-\lambda_1))$. Note that

$$\begin{aligned}
&\hat{\theta}_{\mathrm{cr}}-\frac{1}{2}(\theta_1+\tilde{\theta}_2-\lambda_0-\lambda_1)\\
&\quad=\underbrace{\frac{1}{2}\left\{\frac{1}{n_1}\sum_{i=1}^{n}A_i\left(Y_{i1}^{(1)}-Y_{i2}^{(10)}-\theta_1+\tilde{\tau}+\lambda_0\right)+\frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)\left(Y_{i2}^{(01)}-Y_{i1}^{(0)}-\tilde{\theta}_2-\tilde{\tau}+\lambda_1\right)\right\}}_{M_1}\\
&\quad\underbrace{-\frac{\tilde{\tau}}{2}\cdot\frac{1}{n_1}\sum_{i=1}^{n}A_i+\frac{\tilde{\tau}}{2}\cdot\frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)}_{M_2}
\end{aligned}$$

We can show that

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

$$M_2=0$$

and

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

A.2 Proof of Theorem 2

Theorem 3 is proved from applying Theorem 1 and Corollary 1 in Ye et al. (2022a) with Δ_i being the outcome. From $n^{-1}\sum_{i=1}^n A_i(X_i - \bar{X}) = n^{-1}n_1(\bar{X}_1 - \bar{X}) = O_p(n^{-1/2})$, and $\hat{\beta}_a = \beta_a + o_p(1)$ from Lemma 3 in Ye et al. (2022a), we have that

$$\hat{\theta}_{\text{cr,adj}} = \frac{1}{2}\left\{\frac{1}{n_1}\sum_{i=1}^n A_i\left[Y_{i1} - Y_{i2} - \beta_1^T(X_i - \bar{X})\right] - \frac{1}{n_0}\sum_{i=1}^n (1-A_i)\left[Y_{i1} - Y_{i2} - \beta_0^T(X_i - \bar{X})\right]\right\} + o_p(n^{-1/2})$$

Then,

$$\begin{aligned}
\hat{\theta}_{\text{cr,adj}} - \theta^\star &= \frac{1}{2n_1}\sum_{i=1}^n A_i\{Y_{i1}^{(1)} - Y_{i2}^{(10)} - (\theta_1 - \tilde{\tau} - \lambda_0) - \beta_1^T(X_i - \bar{X})\} \\
&\quad - \frac{1}{2n_0}\sum_{i=1}^n (1-A_i)\{Y_{i1}^{(0)} - Y_{i2}^{(01)} - (\tilde{\theta}_2 + \tilde{\tau} - \lambda_1) - \beta_0^T(X_i - \bar{X})\} + o_p(n^{-1/2}) \\
&= \underbrace{\frac{1}{2n_1}\sum_{i=1}^n A_i\{Y_{i1}^{(1)} - Y_{i2}^{(10)} - (\theta_1 - \tilde{\tau} - \lambda_0) - \beta_1^T(X_i - \mu_X)\}}_{M_{11}} \\
&\quad \underbrace{- \frac{1}{2n_0}\sum_{i=1}^n (1-A_i)\{Y_{i1}^{(0)} - Y_{i2}^{(01)} - (\tilde{\theta}_2 + \tilde{\tau} - \lambda_1) - \beta_0^T(X_i - \mu_X)\}}_{M_{12}} \\
&\quad + \underbrace{\frac{1}{2}(\beta_1 - \beta_0)^T(\bar{X} - \mu_X)}_{M_2} + o_p(n^{-1/2})
\end{aligned}$$

Consider the random vector

$$\sqrt{n}\begin{pmatrix} E_n\left[A_i\{Y_{i1}^{(1)} - Y_{i2}^{(10)} - (\theta_1 - \tilde{\tau} - \lambda_0) - \beta_1^T(X_i - \mu_X)\}\right] \\ E_n\left[(1-A_i)\{Y_{i1}^{(0)} - Y_{i2}^{(01)} + (\tilde{\theta}_2 + \tilde{\tau} - \lambda_1) - \beta_0^T(X_i - \mu_X)\}\right] \\ E_n\left[(X_i - \mu_X)\right] \end{pmatrix}. \tag{9}$$

where $E_n[K_i] := \frac{1}{n}\sum_{i=1}^n K_i$. Since every component in (9) is an average of independent terms. By the assumption that all second moments are finite, the central limit theorem justifies that (9) is asymptotically normal with mean 0 as $n \to \infty$. This implies that $\sqrt{n}(M_{11} + M_{12} + M_2)$ is asymptotically normal.

In what follows, we derive the asymptotic variance of $\hat{\theta}_{\text{cr,adj}} - \theta^\star$. Note that

$$\sqrt{n}(M_{11} - M_{12}) \mid A_1, \ldots, A_n \xrightarrow{d} N\left(0, \frac{\text{Var}(Y_{i1}^{(1)} - Y_{i2}^{(10)} - \beta_1^T X_i)}{4\pi_1} + \frac{\text{Var}(Y_{i1}^{(0)} - Y_{i2}^{(01)} - \beta_0^T X_i)}{4\pi_0}\right),$$

$$\sqrt{n}M_2 \mid A_1, \ldots, A_n \xrightarrow{d} N\left(0, \frac{1}{4}(\beta_1 - \beta_0)^T \text{Var}(X)(\beta_1 - \beta_0)\right)$$

Therefore,

$$\sqrt{n}(\hat{\theta}_{\text{cr,adj}} - \theta^\star) \xrightarrow{d} N(0, \tilde{\sigma}^2_{\text{cr,adj}}),$$

where $\tilde{\sigma}^2_{\text{cr,adj}} = (4\pi_1)^{-1}\text{Var}(Y_{i1}^{(1)} - Y_{i2}^{(10)} - \boldsymbol{\beta}_1^T \boldsymbol{X}_i) + (4\pi_0)^{-1}\text{Var}(Y_{i1}^{(0)} - Y_{i2}^{(01)} - \boldsymbol{\beta}_0^T \boldsymbol{X}_i) + 4^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T \text{Var}(\boldsymbol{X})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$.

Part (b) is a direct consequence of $(4\pi_1)^{-1}\text{Var}(Y_{i1}^{(1)} - Y_{i2}^{(10)} - \beta_1^T X_i) + (4\pi_0)^{-1}\text{Var}(Y_{i2}^{(01)} - Y_{i1}^{(0)} - \beta_0^T X_i) + 4^{-1}(\beta_1 - \beta_0)^T \text{Var}(X)(\beta_1 - \beta_0) = (4\pi_1)^{-1}\text{Var}(Y_{i1}^{(1)} - Y_{i2}^{(10)}) + (4\pi_0)^{-1}\text{Var}(Y_{i2}^{(01)} - Y_{i1}^{(0)}) - 4^{-1}(\beta_1 + \beta_0)^T \text{Var}(X)(\beta_1 + \beta_0)$ from using the definitions of β_0, β_1.

B Controlling for type I error when $\lambda_0 + \lambda_1 < 0$

Similar to Yi et al. (2022), we consider a sensitivity parameter $\Lambda < 0$ that bounds the bias, i.e., $\frac{1}{2}(\lambda_0 + \lambda_1) \geq \Lambda$. Then the rejection region

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

can control the type I error at level α.

For implementation of the sensitivity analysis, practitioners are not required to specify the value of the sensitivity parameter Λ. Results from the sensitivity parameter can be summarized by the "tipping point" – the magnitude of Λ that would be needed such that the null hypothesis can no longer be rejected (Cornfield et al., 1959; Rosenbaum, 2020). If such a value of Λ is deemed implausible, then we still have evidence to reject the null hypothesis.