Research Article

Yanyao Yi#, Ying Zhang#, Yu Du, and Ting Ye*

Testing for treatment effect twice using internal and external controls in clinical trials

https://doi.org/10.1515/jci-2022-0018
received March 08, 2022; accepted August 24, 2023

Abstract: Leveraging external controls – relevant individual patient data under control from external trials or real-world data – has the potential to reduce the cost of randomized controlled trials (RCTs) while increasing the proportion of trial patients given access to novel treatments. However, due to lack of randomization, RCT patients and external controls may differ with respect to covariates that may or may not have been measured. Hence, after controlling for measured covariates, for instance by matching, testing for treatment effect using external controls may still be subject to unmeasured biases. In this article, we propose a sensitivity analysis approach to quantify the magnitude of unmeasured bias that would be needed to alter the study conclusion that presumed no unmeasured biases are introduced by employing external controls. Whether leveraging external controls increases power or not depends on the interplay between sample sizes and the magnitude of treatment effect and unmeasured biases, which may be difficult to anticipate. This motivates a combined testing procedure that performs two highly correlated analyses, one with and one without external controls, with a small correction for multiple testing using the joint distribution of the two test statistics. The combined test provides a new method of sensitivity analysis designed for data fusion problems, which anchors at the unbiased analysis based on RCT only and spends a small proportion of the type I error to also test using the external controls. In this way, if leveraging external controls increases power, the power gain compared to the analysis based on RCT only can be substantial; if not, the power loss is small. The proposed method is evaluated in theory and power calculations, and applied to a real trial.

Keywords: causal inference, data fusion, integrative data analysis, sensitivity analysis

MSC 2020: 62-08, 62D20, 62P10

1 Introduction

1.1 Use of external controls in randomized controlled trials (RCTs)

RCTs are the gold standard for generating high-quality causal evidence of new treatments and have long been recognized as the standard method to support key decisions in the drug development process [1,2]. However, despite its clear advantages, the traditional paradigm of conducting RCTs has been increasingly criticized for

# Yi and Zhang contribute equally.

*** Corresponding author: Ting Ye,** Department of Biostatistics, University of Washington, Seattle, WA 98195, United States, e-mail: tingye1@uw.edu
Yanyao Yi, Ying Zhang, Yu Du: Global Statistical Sciences, Eli Lilly and Company, Indianapolis, IN 46285, United States

failing to meet contemporary needs. In certain settings, for example, in HIV prevention [3,4], oncology [5], and neurology [6], randomizing patients to placebo may be difficult for ethical or feasibility reasons. Moreover, adequately powered RCTs are becoming more and more impractical as a growing number of new treatments are targeted toward rare diseases or biomarker-defined subgroups of patients in the era of precision medicine [7].

Meanwhile, a plethora of real-world data (RWD) have been curated for administrative or research purposes and are becoming accessible to researchers in the form of disease registries, administrative claims databases, and electronic health records. These rich data sources can produce valuable insights, i.e., real-world evidence (RWE), into the effect of treatments in routine, daily practice. However, researchers almost ubiquitously caution against possible bias from unmeasured confounding when using RWD.

Being well aware of the limitations of using either RCT or RWD alone, the idea of using RWD to supplement RCT has gained growing interest in recent years. As forcefully argued in the study of Eichler et al. [7], "the future is not about RCTs vs RWE but RCTs and RWE." There are numerous opportunities in how the integration of RCTs and RWD can achieve fruitful results that using either RCT or RWD alone cannot [8–10]. Among those, an important theme is on augmenting the RCT with RWD to increase efficiency [11–16], and particularly, constructing an externally augmented control arm in the analysis of RCTs [17–20]. Leveraging external controls – relevant individual patient data under control from external trials or RWD – has the potential to reduce the cost of RCTs while increasing the proportion of trial patients given access to novel treatments.

Using external controls is not an entirely new idea. Criteria for evaluating what constitute an acceptable external control arm are proposed in the study of Pocock [21]. It was discussed 20 years ago by the [22, E10 Section 2.5], and also recognized by European Medicines Agency [23], US Food and Drug Administration [24], and National Cancer Institute [25] as one direction to modernize clinical trials. In fact, properly selected external controls (e.g., using propensity score matching) have shown early promise, and several drugs have already been approved based on external control groups [26–28].

Using external controls typically requires the exchangeability condition, i.e., all patient characteristics that affect the potential outcome under control and differ between the trial population and the external control population are measured [29]. While careful adjustment for observed covariates can probably render the exchangeability assumption to hold approximately, the analysis may still be biased due to unmeasured covariates related to "difficulties in reliably selecting a comparable population because of potential changes in medical practice, lack of standardized diagnostic criteria or equivalent outcome measures, and variability in follow-up procedures" [24]. To reduce the potential biases from using external controls, an intuitive frequentist approach is "test-then-pool" that first tests for the comparability of the external controls and internal controls before leveraging external controls [20]. Bayesian methods that rely on power priors have also been popular, which use the likelihood of the external data to a specified power as the prior distribution [30,31]. As such, one can use power priors to adjust the weight allocated to the external information according to the levels of comparability between the external control and the internal data. However, these methods lack formal statistical theory on how the unmeasured biases might affect the validity and efficiency of the proposed procedures.

In this article, we take a different perspective to this problem and propose a sensitivity analysis approach to quantify the magnitude of unmeasured bias that would be needed to alter the study conclusion that presumed no unmeasured biases are introduced by employing external controls [32]. With the unbiased RCT-only test as the benchmark, leveraging external controls increases power or not depends on the interplay between sample sizes and the magnitude of treatment effect and unmeasured biases, which may be difficult to anticipate. This motivates a combined testing procedure that performs both tests, one with and one without external controls, correcting for multiple testing using the joint distribution of the two test statistics. Because the two tests are highly correlated, this correction for multiple testing is small. Interestingly, the proposed combined testing procedure can be viewed as a new method of sensitivity analysis designed for data fusion problems that anchors at the unbiased analysis based on RCT only and "spends" a small proportion of the type I error (i.e., the cost of multiple testing) to also test using the pooled controls. In this way, if leveraging external controls increases power, the power gain compared to the RCT-only test can be substantial; if not, the power loss is small. Before introducing technical details, it is useful to consider a motivating example.

1.2 Example: an RCT in patients with type-2 diabetes

Consider a non-inferiority, phase 3 RCT (referred to as the internal trial, ClinicalTrials.gov number, NCT01894568) comparing a new basal insulin, insulin peglispro, to insulin glargine as the control in Asian insulin-naïve patients with type-2 diabetes using a noninferiority margin of 0.4% [33]. The primary endpoint is the change in hemoglobin A1c (HbA1c) from baseline to 26 weeks of treatment. HbA1c is a continuous-valued measure of average blood glucose in the past 3 months. Before this trial, a phase 3 RCT of similar design (referred to as the external trial, ClinicalTrials.gov number, NCT01435616) has been conducted in the North America and Europe [34], whose control arm will be used as the source of external controls.

We focus on the overweight and obese population, which are, respectively, defined as 23≤ body mass index (BMI) <25 and BMI $\geq$ 25 for the internal trial according to the Asia-Pacific guidelines, and 25 $\leq$ BMI $<$ 30 and BMI $\geq$ 30 for the external trial according to the World Health Organization classifications [35]. There are in total 159 patients under treatment and 150 patients under control in the internal RCT, and 486 patients under control in the external trial. We match 159 similar external controls to the 159 treated patients in the internal RCT using optimal matching based on a robust Mahalanobis distance and a caliper on the propensity score. See Rosenbaum [36, Part II] for discussion of these matching techniques. Table 1 describes covariate balance in the 159 matched pairs. All variables have standardized differences less than 0.13 and are considered sufficiently balanced [37].

Using only the internal RCT, 159 patients under treatment and 150 under control, we conduct a Z-test with the noninferiority margin of 0.4% and obtain a one-sided p-value 7.92×10^{-7}. In this analysis, the evidence that

Table 1: Covariate balance after matching in 159 matched pairs of one treated patient in the RCT and one external control patient

	Treated (n_1 = 159)	External control (n_e = 159)	Standardized mean difference
Age (years)	57.45	57.16	0.03
Female (fr)	0.41	0.41	0.00
Overweight (fr)	0.28	0.28	0.00
Obese (fr)	0.72	0.72	0.00
Diabetes duration (years)	12.08	11.74	0.05
Hypertension (fr)	0.66	0.69	−0.07
History of MI (fr)	0.04	0.01	0.13
History of CR (fr)	0.04	0.02	0.13
History of CABG (fr)	0.01	0.00	0.05
Lipid lowering medication (fr)	0.61	0.57	0.09
Statin Use (fr)	0.51	0.50	0.01
Non-statin lipid lowering medication (fr)	0.15	0.11	0.10
Fasting serum glucose (mg/dL)	164.99	166.25	−0.03
Triglycerides (mg/dL)	139.86	140.06	−0.00
Total cholesterol (mg/dL)	178.92	180.93	−0.05
LDL (mg/dL)	101.42	103.29	−0.06
HDL (mg/dL)	50.41	50.01	0.03
Alanine aminotransferase (U/L)	33.60	33.03	0.03
Aspartate aminotransferase (U/L)	26.99	26.08	0.08
Total bilirubin (mg/dL)	0.58	0.54	0.12
eGFR (mL/min/1.73 m^2)	90.52	87.70	0.13
Baseline sulfonylureas or meglitinides use (fr)	0.86	0.86	0.02
Smoking (fr)	0.46	0.43	0.06
Baseline HbA1c (%)	8.57	8.59	0.03

Abbreviations: CABG = coronary artery bypass graft; CR = coronary revascularization; eGFR = estimated glomerular filtration rate based on the modified Modification of Diet in Renal Disease equation; fr = fraction; HbA1c = hemoglobin A1c; HDL = high-density lipoprotein cholesterol; LDL = low-density lipoprotein cholesterol; MI = myocardial infarction.

the new insulin treatment is noninferior to insulin glargine is strong enough when only using the internal controls. On the other hand, under the exchangeability assumption, which implies that the 159 matched external controls are comparable to patients in the internal RCT, we construct an augmented control arm of 309 patients in total and obtain a one-sided p-value 1.88×10^{-7}. Again, we find strong evidence of noninferiority; however, an investigator may be in doubt about the exchangeability assumption due to the influence of regions on the outcome. Then a natural question is could the one-sided p-value of 1.88×10^{-7} be due to regions rather than the effect of treatment? If the study conclusion from using external controls can be altered by a plausible effect of regions and because the RCT-only test is already powerful enough, the RCT-only test would be a better choice. However, it would be difficult to know this before examining the data. Motivated by the advice of performing multiple analyses with an appropriate correction for multiple testing given by Rosenbaum [38], we propose a combined testing procedure that performs both analyses, controlling for multiple testing using the joint distribution of the two test statistics. In this article, we will demonstrate that the combined test avoids making an inapt choice about whether to use external controls or not, and only has a small loss of power compared to knowing a priori which is the better choice.

1.3 Outline

Section 2 presents a test that uses only the internal controls and another test that also leverages the external controls, and discusses controlling type I error and comparing power without the exchangeability assumption. Section 3 proposes a combined test that performs both tests and studies in detail its statistical properties. Section 4 presents power calculations. Section 5 returns to the real data applications. Section 6 concludes with a discussion.

2 Testing using internal and external controls

2.1 Testing under exchangeability

There is an RCT denoted as $D = 1$. Let A be a binary treatment, where $A = 1$ denotes treatment and $A = 0$ denotes control, X a vector of observed baseline covariates, $Y^{(a)}$ the potential outcome under $A = a$, for $a = 0, 1$. Throughout the article, we assume consistency and Stable Unit Treatment Value Assumption (SUTVA) so that the observed outcome satisfies $Y = AY^{(1)} + (1 - A)Y^{(0)}$ [39]. Our estimand of interest is the average treatment effect in the RCT population $\theta^{\star} = E(Y^{(1)}|D = 1) - E(Y^{(0)}|D = 1)$. In particular, we consider testing a one-sided hypothesis:

$$H_0 : \theta^{\star} = \theta_0 \quad \text{versus } H_A : \theta^{\star} > \theta_0.$$

The other direction can be considered in the same way. Combining both one-sided tests and applying Bonferroni correction give a two-sided test [40, Section 4.2], and by inversion, a confidence interval.

Write the RCT sample as $(Y_i, X_i, A_i, D_i = 1),\ i = 1, \ldots, n_r$, which is assumed to be independent and identically distributed according to the joint law of $(Y^{(1)}, Y^{(0)}, X, A)|D = 1$. Randomization in the RCT guarantees that $A \perp (Y^{(1)}, Y^{(0)}, X)|D = 1$ and $P(A = a|D = 1) = \pi_a > 0$ for $a = 0, 1$, with π_a known and $\pi_0 + \pi_1 = 1$. Let $\overline{Y}_a$ and S_a^2, respectively, be the sample mean and sample variance of the responses Y_i's from RCT subjects under treatment a, for $a = 0, 1$. Hence, the null hypothesis H_0 can be tested using a simple Z-statistic:

$$T_1 = \frac{\overline{Y}_1 - \overline{Y}_0 - \theta_0}{\sqrt{n_1^{-1}S_1^2 + n_0^{-1}S_0^2}},$$

where n_1 and n_0 are, respectively, the number of RCT patients under treatment and control. Based on T_1, we reject H_0 when $T_1 \geq z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1 - \alpha)$th quantile of the standard normal distribution.

To supplement the RCT using external controls, one approach is to first extract external data for patients under control based on the inclusion/exclusion criteria of the RCT and then proceed by matching these external patients to the RCT patients based on their similarity in observed baseline information X [27]. Let $D = 0$ denote the matched external controls, and thus $D = 0$ implies $A = 0$. Write the matched external controls as $(Y_i, X_i, A_i = 0, D_i = 0)$, $i = 1, \ldots, n_e$, which is assumed to be independent and identically distributed according to the joint law of $(Y^{(0)}, X)|A = 0, D = 0$. Suppose that matching has rendered the baseline observed covariates comparable between the RCT and external controls, i.e., $D \perp X$, and that these baseline covariates X explain all differences between the RCT and external controls, i.e., the exchangeability assumption $D \perp Y^{(0)}|X$ holds. This implies $D \perp (Y^{(0)}, X)$ and thus $E(Y^{(0)}|D = 1) = E(Y^{(0)}|D = 0)$. Let $\overline{Y}_e$ be the sample mean of the responses Y_i's from the external controls, and $w\overline{Y}_0 + (1 - w)\overline{Y}_e$ be a weighted average of mean responses for the two control groups, where $w \in [0, 1]$ is a pre-specified weight, which could reflect the proportion of the internal control in the two control groups combined. Therefore, the null hypothesis H_0 can also be tested borrowing information from the external controls using

$$T_2(w) = \frac{\overline{Y}_1 - \{w\overline{Y}_0 + (1 - w)\overline{Y}_e\} - \theta_0}{\sqrt{n_1^{-1}S_1^2 + w^2 n_0^{-1}S_0^2 + (1 - w)^2 n_e^{-1}S_e^2}},$$

where S_e^2 is the sample variance of the responses Y_i's from external controls. We make two remarks about $T_2(w)$. First, $T_2(w)$ is constructed assuming independence between the RCT and external controls, which means that $T_2(w)$ may be conservative due to correlation induced by matching [41] but usually to a small extent as the correlation is typically small [42]. Second, $T_2(w)$, $w \in [0, 1]$ defines a family of statistics that includes $T_2(1) = T_1$ as a special case. Among those, the exchangeability assumption implies the optimal w that maximizes the efficiency of $T_2(w)$ is proportional to the sample size, i.e., the optimal w equals $(n_r\pi_0)/(n_r\pi_0 + n_e)$. One can also choose different values of w to reflect the weights allocated to the two control groups.

2.2 Controlling type I error without exchangeability

The aforementioned approach of leveraging external controls relies on the exchangeability assumption, which may not hold because the RCT patients and external controls may differ with respect to covariates that may not have been measured. Without exchangeability, $\overline{Y}_1 - \{w\overline{Y}_0 + (1 - w)\overline{Y}_e\}$ is not necessarily centered at θ_0 under H_0 and rejecting the null hypothesis when $T_2(w) \geq z_{1-\alpha}$ may inflate type I error.

Define $\Delta^* = E(Y^{(0)}|D = 1) - E(Y^{(0)}|D = 0)$, which may be nonzero when exchangeability does not hold. This could occur, for example, if an important prognostic variable is unobserved and left uncontrolled, or if a variable that differs in distribution between $D = 0$ and $D = 1$ (such as region) cannot be matched. The correct rejection region for a size-α test based on $T_2(w)$ is

$$T_2(w) - \frac{(1 - w)\Delta^*}{\sqrt{n_1^{-1}S_1^2 + w^2 n_0^{-1}S_0^2 + (1 - w)^2 n_e^{-1}S_e^2}} > z_{1-\alpha},$$

which is infeasible because Δ^* is unknown. To deal with this issue, a tempting choice is to estimate Δ^* and "debiase" the numerator of $T_2(w)$ to make it mean zero. Nonetheless, estimating Δ^* introduces additional variation, which negatively affects the efficiency of the test. In particular, if one estimates Δ^* by $\overline{Y}_0 - \overline{Y}_e$, the debiased numerator of $T_2(w)$ becomes $\overline{Y}_1 - \overline{Y}_0 - \theta_0$, and the resulting test statistic (after appropriately adjusting for its denominator to reflect the variability of the numerator) becomes equivalent to T_1, the test statistic without using any external controls.

In order to borrow information from external controls while still controlling type I error, we consider departures from the exchangeability through the lens of a sensitivity analysis [36]. Specifically, we consider a sensitivity parameter Δ_0 such that it bounds the magnitude of bias Δ^*, i.e., $\Delta_0 \geq \Delta^*$. Define

$$T_{2,\Delta_0}(w) = \frac{\overline{Y}_1 - \{w\overline{Y}_0 + (1 - w)\overline{Y}_e\} - \theta_0 - (1 - w)\Delta_0}{\sqrt{(n_r\pi_1)^{-1}S_1^2 + w^2(n_r\pi_0)^{-1}S_0^2 + (1 - w)^2 n_e^{-1}S_e^2}}.$$

Because $\Delta^\star \le \Delta_0$, the reject region $T_{2,\Delta_0}(w) \ge z_{1-\alpha}$ controls type I error at level α. As a special case when $\Delta^\star \le \Delta_0$ holds with $\Delta_0 = 0$ (e.g., under exchangeability), $T_{2,\Delta_0}(w) \ge z_{1-\alpha}$ becomes $T_2(w) \ge z_{1-\alpha}$, the reject region under exchangeability. As Δ_0 increases, there is greater uncertainty about how the exchangeability might be violated, leading to more stringent rejection criterion to control type I error. The reject region $T_{2,\Delta_0}(w) \ge z_{1-\alpha}$ is sharp under $\Delta^\star \le \Delta_0$ in the sense that they are of size-α when $\Delta^\star = \Delta_0$, so it cannot be improved unless further information is provided.

2.3 Power comparison without exchangeability

Write $\sigma_a^2 = \operatorname{Var}(Y^{(a)}|D = 1)$, for $a = 0, 1$, and $\sigma_e^2 = \operatorname{Var}(Y^{(0)}|D = 0)$. Under the alternative hypothesis $H_A : \theta^\star > \theta_0$, the power of T_1 is the probability of event $T_1 \ge z_{1-\alpha}$, which is asymptotically equal to

$$1 - \Phi\left(z_{1-\alpha} + \frac{\sqrt{n_r}(\theta_0 - \theta^\star)}{\sqrt{\pi_1^{-1}\sigma_1^2 + \pi_0^{-1}\sigma_0^2}} \right), \tag{1}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution. In parallel, the power of $T_{2,\Delta_0}(w)$ is the probability of event $T_{2,\Delta_0}(w) \ge z_{1-\alpha}$, which is asymptotically equal to

$$1 - \Phi\left(z_{1-\alpha} + \frac{\sqrt{n_r}(\theta_0 - \theta^\star) + \sqrt{n_r}(1-w)(\Delta_0 - \Delta^\star)}{\sqrt{\pi_1^{-1}\sigma_1^2 + w^2\pi_0^{-1}\sigma_0^2 + (1-w)^2 n_r n_e^{-1}\sigma_e^2}} \right). \tag{2}$$

Several remarks are in order based on the above power formulas. First, the power of $T_{2,\Delta_0}(w)$ is larger than that of T_1 if and only if

$$\frac{\theta_0 - \theta^\star + (1-w)(\Delta_0 - \Delta^\star)}{\sqrt{\pi_1^{-1}\sigma_1^2 + w^2\pi_0^{-1}\sigma_0^2 + (1-w)^2 n_r n_e^{-1}\sigma_e^2}} \le \frac{\theta_0 - \theta^\star}{\sqrt{\pi_1^{-1}\sigma_1^2 + \pi_0^{-1}\sigma_0^2}}.$$

For instance, when $\Delta_0 = \Delta^\star$, i.e., the specified upper bound for $\Delta^\star$ is tight, and $\sigma_0^2 = \sigma_e^2$, i.e., the variance of Y for the two control groups are equal, simple algebra reveals that the power of $T_{2,\Delta_0}(w)$ is always larger than that of T_1 for any w satisfying $\max(0, (n_r\pi_0 - n_e)/(n_r\pi_0 + n_e)) \le w < 1$.

Second, we can derive the oracle w that maximizes the power of $T_{2,\Delta_0}(w)$. Let $\kappa = (\pi_0^{-1}\sigma_0^2)/(\pi_1^{-1}\sigma_1^2 + \pi_0^{-1}\sigma_0^2)$, the optimal w takes the following form:

$$w_{\text{opt}} = \begin{cases} 1, & \text{when } \Delta_0 - \Delta^\star \ge \kappa(\theta^\star - \theta_0) > 0, \\ 1 - \dfrac{(\Delta_0 - \Delta^\star)(\pi_1^{-1}\sigma_1^2 + \pi_0^{-1}\sigma_0^2) + (\theta_0 - \theta^\star)\pi_0^{-1}\sigma_0^2}{(\theta_0 - \theta^\star)(n_r n_e^{-1}\sigma_e^2 + \pi_0^{-1}\sigma_0^2) + (\Delta_0 - \Delta^\star)\pi_0^{-1}\sigma_0^2}, & \text{when } \kappa(\theta^\star - \theta_0) > \Delta_0 - \Delta^\star \ge 0, \end{cases} \tag{3}$$

where the first case is when Δ_0 is specified too large, the power of $T_{2,\Delta_0}(w)$ is maximized at $w = 1$, which means that using the external controls does not lead to efficiency gain. As an illustration, under the special case that $\pi_1^{-1}\sigma_1^2 = \pi_0^{-1}\sigma_0^2 = n_r n_e^{-1}\sigma_e^2$, when $\Delta_0 - \Delta^\star > (\theta^\star - \theta_0)/2$, the optimal w is 1, whereas when $(\theta^\star - \theta_0)/2 > \Delta_0 - \Delta^\star > 0$, the optimal w is $1 - \{(\theta_0 - \theta^\star) + 2(\Delta_0 - \Delta^\star)\}/\{2(\theta_0 - \theta^\star) + (\Delta_0 - \Delta^\star)\}$. Under another special case when $\Delta^\star = \Delta_0$ and $\sigma_1 = \sigma_0 = \sigma_e$, w_{opt} becomes $(n_r\pi_0)/(n_r\pi_0 + n_e)$, which agrees with the optimal w under exchangeability discussed in Section 2.1. The proof of (3) is given in the supplementary material.

Finally, we compare the two tests T_1 and $T_{2,\Delta_0}(w)$ in terms of their limiting power as the sample sizes grow to infinity. When $\theta^\star > \theta_0$ and $\lim_{n_r \to +\infty}\sqrt{n_r}(\theta^\star - \theta_0) = +\infty$ (e.g., when $\theta^\star$, θ_0 are two constants), then the power of T_1 goes to 1 as $n_r \to \infty$. In contrast, the limiting power of $T_{2,\Delta_0}(w)$ depends on specifications of w and Δ_0. Specifically, as can be seen from the power formula in (2), when $\min(n_r, n_e) \to +\infty$, the power of $T_{2,\Delta_0}(w)$ tends to 1 if $\theta_0 - \theta^\star + (1-w)(\Delta_0 - \Delta^\star) < 0$, i.e., when $\Delta_0 < (\theta^\star - \theta_0)/(1-w) + \Delta^\star$, and to 0 if $\theta_0 - \theta^\star + (1-w)(\Delta_0 - \Delta^\star) > 0$, i.e., when $\Delta_0 > (\theta^\star - \theta_0)/(1-w) + \Delta^\star$. In other words, there exists a w-dependent number $\widetilde{\Delta}(w) = (\theta^\star - \theta_0)/(1-w) + \Delta^\star$ that characterizes the limiting behavior of $T_{2,\Delta_0}(w)$ under the alternative: the power of

$T_{2,\Delta_0}(w)$ tends to 1 if $\Delta_0 < \widetilde{\Delta}(w)$ and to 0 if $\Delta_0 > \widetilde{\Delta}(w)$ as $\min(n_r, n_e) \to +\infty$. In plain language, since the maximum bias Δ_0 counteracts with $\theta^\star - \theta_0$, the difference we would like to detect, when Δ_0 exceeds a certain threshold $\widetilde{\Delta}(w)$, the maximum bias starts to dominate the true difference $\theta^\star - \theta_0$, resulting in no power to detect the difference. This number $\widetilde{\Delta}(w)$ is analogous to the design sensitivity in the literature of sensitivity analysis [36,43].

3 A combined test

Should we leverage external controls? In other words, is it better to use the test statistic T_1 constructed solely based on the RCT or the test statistic $T_{2,\Delta_0}(w)$ that additionally leverages the external controls? We know from the above theory and analysis that the answer to this question depends upon the context, specifically upon the nature and size of the treatment effect, and the specification of w and Δ_0, that might be difficult to anticipate prior to examining the data. As Motivated in Section 1, we propose a combined testing procedure that performs both T_1 and $T_{2,\Delta_0}(w)$, correcting for multiple testing using the joint distribution of the two test statistics.

Under H_0, the joint distribution of $(T_1, T_{2,\Delta^\star}(w))$ is asymptotically bivariate normal, satisfying

$$\begin{pmatrix} T_1 \\ T_{2,\Delta^\star}(w) \end{pmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right),$$

where

$$\rho = \frac{\pi_1^{-1}\sigma_1^2 + w\pi_0^{-1}\sigma_0^2}{\sqrt{(\pi_1^{-1}\sigma_1^2 + \pi_0^{-1}\sigma_0^2)(\pi_1^{-1}\sigma_1^2 + w^2\pi_0^{-1}\sigma_0^2 + (1-w)^2 n_r n_e^{-1}\sigma_e^2)}}.$$

Again, for illustration, consider the special case that $\pi_1^{-1}\sigma_1^2 = \pi_0^{-1}\sigma_0^2 = n_r n_e^{-1}\sigma_e^2$, then ρ increases as w increases from 0 to 1, and thus ρ ranges between 0.5 and 1.

Consider the testing procedure that, for any specified Δ_0 and w, rejects H_0 if

$$\max(T_1, T_{2,\Delta_0}(w)) \geq c_{1-\alpha;\,\rho}, \tag{4}$$

where $c_{1-\alpha;\,\rho}$ satisfies $\Phi_{2,\rho}(c_{1-\alpha;\,\rho}) = 1 - \alpha$, $\Phi_{2,\rho}(x, y)$ is the probability of the two-dimensional lower orthant $(-\infty, x] \times (-\infty, y]$ for a bivariate normal distribution with expectation $(0, 0)^T$, unit variances, and correlation coefficient ρ, and write $\Phi_{2,\rho}(x) = \Phi_{2,\rho}(x, x)$. This combined testing procedure is able to control the type I error for any $\Delta^\star \in [-\infty, \Delta_0]$ because

$$P_{H_0}(\max(T_1, T_{2,\Delta_0}(w)) \geq c_{1-\alpha;\,\rho}) \leq P_{H_0,\Delta^\star=\Delta_0}(\max(T_1, T_{2,\Delta^\star}(w)) \geq c_{1-\alpha;\,\rho}) = \alpha.$$

In what follows, we establish several attractive features of the combined test. Note that under the alternative hypothesis, the power of the combined test – the probability of event (4) – is

$$\begin{aligned} &P(\max(T_1, T_{2,\Delta_0}(w)) \geq c_{1-\alpha;\,\rho}) \\ &\approx 1 - \Phi_{2,\rho}\left(c_{1-\alpha;\,\rho} + \underbrace{\frac{\sqrt{n_r}(\theta_0 - \theta^\star)}{\sqrt{\pi_1^{-1}\sigma_1^2 + \pi_0^{-1}\sigma_0^2}}}_{B_1},\quad c_{1-\alpha;\,\rho} + \underbrace{\frac{\sqrt{n_r}(\theta_0 - \theta^\star) + \sqrt{n_r}(1-w)(\Delta_0 - \Delta^\star)}{\sqrt{\pi_1^{-1}\sigma_1^2 + w^2\pi_0^{-1}\sigma_0^2 + (1-w)^2 n_r n_e^{-1}\sigma_e^2}}}_{B_2} \right), \end{aligned} \tag{5}$$

where ≈ means asymptotic approximation. This leads to the first observation that the power of the combined test is generally larger than the worst of the two component tests, i.e., $\text{Power}_c \geq \min(\text{Power}_1, \text{Power}_2)$, where Power_1, Power_2, Power_c are, respectively, the asymptotic power of T_1, $T_{2,\Delta_0}(w)$, and the combined test. This can be seen from noting that

$$\begin{aligned}1 - \text{Power}_c &= \Phi_{2,\rho}(c_{1-\alpha;\,\rho} + B_1, c_{1-\alpha;\,\rho} + B_2)\\ &\le \Phi_{2,\rho}(c_{1-\alpha;\,\rho} + \min(B_1, B_2), +\infty)\\ &= \Phi(c_{1-\alpha;\,\rho} + \min(B_1, B_2))\\ &= \Phi(z_{1-\alpha} + \max(B_1, B_2) - \{|B_1 - B_2| - (c_{1-\alpha;\,\rho} - z_{1-\alpha})\})\\ &\le \Phi(z_{1-\alpha} + \max(B_1, B_2))\\ &= \max\{\Phi(z_{1-\alpha} + B_1), \Phi(z_{1-\alpha} + B_2)\}\\ &= 1 - \min(\text{Power}_1, \text{Power}_2),\end{aligned}$$

where the second inequality holds when $|B_1 - B_2| \ge (c_{1-\alpha;\,\rho} - z_{1-\alpha})$, i.e., when the power of the two component tests are not too similar.

Moreover, not only is the power of the combined test better than the worst of the two component tests in finite sample, it is also close to the better of the two component tests in finite sample, and equal to the better of the two component tests in the limit. To see this, we bound the difference in power as follows:

$$\begin{aligned}\max(\text{Power}_1, \text{Power}_2) - \text{Power}_c &= \Phi_{2,\rho}(c_{1-\alpha;\,\rho} + B_1, c_{1-\alpha;\,\rho} + B_2) - \Phi(z_{1-\alpha} + \min(B_1, B_2))\\ &\le \Phi(c_{1-\alpha;\,\rho} + \min(B_1, B_2)) - \Phi(z_{1-\alpha} + \min(B_1, B_2))\\ &\le 1 - 2\Phi((z_{1-\alpha} - c_{1-\alpha;\,\rho})/2).\end{aligned}$$

It is helpful to anchor several values of $c_{1-\alpha;\,\rho}$ and the upper bound $1 - 2\Phi((z_{1-\alpha} - c_{1-\alpha;\,\rho})/2)$ in terms of different α and ρ. When $\alpha = 0.025$ and for $\rho = 0.5, 0.7, 1$, the critical values are $c_{1-\alpha;0.5} = 2.21$, $c_{1-\alpha;0.7} = 2.18$, $c_{1-\alpha;1} = 1.96$, and respectively, the upper bounds are 0.100, 0.088, 0. This means that because of the high correlation between T_1 and $T_{2,\Delta^*}(w)$, the price paid for multiple testing is generally small. With regard to the limiting power, it is also easy to see that for fixed θ_0 and $\theta^* > \theta_0$, $B_1 \to -\infty$ as the sample size n_r increases. Hence, the combined test always has its power approaching 1 as $n_r \to \infty$, just like the test T_1 that only uses RCT data, which is not the case for $T_{2,\Delta^*}(w)$ as discussed in Section 2.3. This further shows the advantage of the combined test.

For implementation of the sensitivity analysis (either T_{2,Δ_0} or the combined test), practitioners are not required to specify the value of the sensitivity parameter Δ_0. Following the pioneering work by Cornfield et al. [44] and the sensitivity analysis literature [36], results from the combined test can be summarized by the "tipping point" – the magnitude of Δ_0 that would be needed such that the null hypothesis can no longer be rejected. If such a value of Δ_0 is deemed implausible, then we still have evidence to reject the null hypothesis based on the combined test. In Section 5, we illustrate the method using a real example.

4 Power calculations

We investigate three factors when conducting power calculations. The first factor concerns the true treatment effect $\theta^* = 0.2$, 0.3, and 0.4. The second factor is the specified value of the maximum bias $\Delta_0 = 0.2, 0.3, 0.4, 0.6$. The third factor is the sample size $n_1 = 50, 100, 150, 200$, with $n_1 : n_0 : n_e = 2 : 1 : 3$. Additional parameters are $\theta_0 = 0$, $\Delta^* = 0.2$, $\sigma_1 = \sigma_0 = \sigma_e = 1$, and $\alpha = 0.025$.

Table 2 summarizes the power of T_1, $T_{2,\Delta_0}(w)$ and the combined test $T_{c,\Delta_0}(w) = \max(T_1, T_{2,\Delta_0}(w))$, calculated, respectively, using (1), (2), and (5). For $T_{2,\Delta_0}(w)$ and $T_{c,\Delta_0}(w)$, we consider two choices of w: the oracle w in (3) that maximizes the power (denoted as w_{opt}), and its value under exchangeability $n_0/(n_0 + n_e) = 1/4$. In the supplementary material, we check powers by simulation, finding good agreement. In the supplementary material, we also include a check of the type I error, which are all close to or below the nominal level, indicating validity of all the tests. In contrast, a naive combined test without correcting for multiple testing cannot control the type I error.

The following is a summary of results in Table 2.

- Across all scenarios, the power of the combined test $T_{c,\Delta_0}(1/4)$ is larger than the worst of the power of T_1 and $T_{2,\Delta_0}(1/4)$, and close to the best of the power of T_1 and $T_{2,\Delta_0}(1/4)$. This supports our theory in Section 3.

- For T_1, its power is not affected by $\varDelta_0$. For $T_{2,\varDelta_0}(1/4)$, its power is mostly larger than that of T_1 when $\varDelta_0 = 0.2, 0.3$, but quickly diminishes as $\varDelta_0$ increases and becomes substantially smaller than that of T_1 when $\varDelta_0 = 0.4, 0.6$ across most scenarios. In comparison, when $\theta^\star = 0.2, 0.3$, the sensitivity parameter $\varDelta_0$ can be as large as 0.3 before the combined test $T_{c,\varDelta_0}(1/4)$ starts to lose power compared to T_1; when $\theta^\star = 0.4$, the sensitivity parameter $\varDelta_0$ can be as large as 0.4. If a $\varDelta_0$ larger than 0.3 or 0.4 is deemed implausible by practitioners, the combined test $T_{c,\varDelta_0}(1/4)$ will have power gain compared to T_1. On the other hand, because the combined test $T_{c,\varDelta_0}(1/4)$ still performs T_1 as one of its component (i.e., anchors at T_1) but with a small adjustment for testing twice, the potential power loss compared to T_1 is never too large. This clearly demonstrates the key advantage of the combined test.
- As the sample size increases, the power of T_1 and $T_{c,\varDelta_0}(1/4)$ always increases. However, as the sample size increases, the power of $T_{2,\varDelta_0}(1/4)$ tends to 0 when $\theta^\star = 0.2$ and $\varDelta_0 = 0.6$, stays unchanged when $\theta^\star = 0.3$ and $\varDelta_0 = 0.6$, and tends to 1 in other cases. This behavior of $T_{2,\varDelta_0}(1/4)$ supports the result that the power of $T_{2,\varDelta_0}(1/4)$ tends to 1 when $\varDelta_0 < \tilde{\varDelta}(1/4)$ and to 0 when $\varDelta_0 > \tilde{\varDelta}(1/4)$ as the sample size increases, where $\tilde{\varDelta}(1/4)$ defined in Section 2.3 equals $4\theta^\star/3 + 0.2$, which is 0.47, 0.60, and 0.73 for $\theta^\star = 0.2, 0.3, 0.4$, respectively.
- Finally, the oracle tests $T_{2,\varDelta_0}(w_{\text{opt}})$ and $T_{c,\varDelta_0}(w_{\text{opt}})$ are included as a reference. The test $T_{2,\varDelta_0}(w_{\text{opt}})$ is more powerful than T_1 and $T_{2,\varDelta_0}(1/4)$, which agrees with our theory as $T_{2,\varDelta_0}(w_{\text{opt}})$ maximizes power among a family of test statistics $\{T_2(w), w \in [0,1]\}$. Observing that the power of $T_{c,\varDelta_0}(1/4)$ and $T_{c,\varDelta_0}(w_{\text{opt}})$ are similar indicates that setting $w = n_0/(n_0 + n_e)$ usually leads to desirable power performance.

5 Application

We revisit the example introduced in Section 1.2 and illustrate how the proposed methods can be applied. Formally, we test the hypothesis that $H_0 : \theta^\star = \theta_0$ versus $H_A : \theta^\star < \theta_0$, with $\theta_0 = 0.4$, which can be equivalently implemented using the tests described in Sections 2 and 3 with Y_i's replaced by $-Y_i$'s and θ_0 replaced by $-\theta_0$. We set the significance level $\alpha = 0.025$.

Using only the internal RCT, $T_1 = 4.80$ with p-value 7.92×10^{-7}, based on which we reject the null hypothesis H_0. This result is solely based on internal controls and thus is invariant to the value of $\varDelta_0$.

Leveraging external controls and let $w = n_0/(n_0 + n_e) = 0.485$, $T_2(w) = 5.08$ with p-value 1.88×10^{-7} when $\varDelta_0 = 0$. Therefore, under the exchangeability assumption, we can also reject the null hypothesis H_0. To gauge the robustness of this conclusion to violation of the exchangeability, we apply the proposed sensitivity analysis. As discussed at the end of Section 3, results of our sensitivity analysis can be summarized by the "tipping point" – the magnitude of $\varDelta_0$ that would be needed such that the null hypothesis can no longer be rejected. In this example, as $\varDelta_0$ increases, the adjusted p-value associated with $T_{2,\varDelta_0}(w)$ increases but remains below $\alpha = 0.025$ for any $\varDelta_0 \leq 0.62$. Namely, two patients with the same observed characteristics (as listed in Table 1), one in the internal RCT and the other in the external trial, may differ in their expected potential outcome under control by up to 0.62, under which the adjusted p-value is still below the significance level α. This means that the significant effect we observe cannot be explained away by unmeasured biases of magnitude up to $\varDelta_0 = 0.62$. If such a large unmeasured bias is deemed implausible, then there is no real doubt that the rejection based on $T_{2,\varDelta_0}$ provides evidence of noninferiority.

Finally, using the combined test, $\max(T_1, T_{2,\varDelta_0}(w)) = 5.08$ with adjusted p-value 3.41×10^{-7} when $\varDelta_0 = 0$. As $\varDelta_0$ increases, the adjusted p-value for the combined test increases but plateaus at 1.41×10^{-6} when $T_1 \geq T_{2,\varDelta_0}(w)$. This means that rejection based on the combined test is insensitive to any value of $\varDelta_0$, i.e., similar to T_1 that only uses the internal RCT, rejection based on the combined test is insensitive to any violation of the exchangeability assumption.

It is also interesting to see the relative performance of T_1, $T_{2,\varDelta_0}(w)$, $T_{c,\varDelta_0}(w)$ when the internal RCT is underpowered, and thus the combined test may be more useful. For this purpose, we randomly sample with replacement 100 patients from the internal RCT, with a target ratio of 4/5 from the treated arm and 1/5 from the control arm. Then T_1 is computed using this subsample from the RCT, while $T_{2,\varDelta_0}(w)$ and $T_{c,\varDelta_0}(w)$ additionally use the external controls that were matched to the sampled treated patients with $w = n_0/(n_0 + n_e)$

Table 2: Theoretical power (in %) for T_1, $T_{2,\Delta_0}(w)$ and the combined test $T_{c,\Delta_0}(w)$ with $w = 1/4$ or w_{opt}, where $\theta_0 = 0$, $\Delta^* = 0.2$, $n_1 : n_0 : n_e = 2 : 1 : 3$, $\sigma_1 = \sigma_0 = \sigma_e = 1$, and $\alpha = 2.5\%$

		$\theta^* = 0.2$					$\theta^* = 0.3$					$\theta^* = 0.4$				
Δ_0	n_1	T_1	$T_2(1/4)$	$T_2(w_{\text{opt}})$	$T_c(1/4)$	$T_c(w_{\text{opt}})$	T_1	$T_2(1/4)$	$T_2(w_{\text{opt}})$	$T_c(1/4)$	$T_c(w_{\text{opt}})$	T_1	$T_2(1/4)$	$T_2(w_{\text{opt}})$	$T_c(1/4)$	$T_c(w_{\text{opt}})$
0.2	50	12.6	21.0	21.0	18.5	18.5	23.1	41.0	41.0	36.5	36.5	37.2	63.7	63.7	58.4	58.4
	100	21.0	37.2	37.2	33.0	33.0	41.0	68.8	68.8	63.7	63.7	63.7	90.4	90.4	87.5	87.5
	150	29.3	51.6	51.6	46.5	46.5	56.4	85.1	85.1	81.3	81.3	80.7	97.9	97.9	97.0	97.0
	200	37.2	63.7	63.7	58.4	58.4	68.8	93.4	93.4	91.1	91.1	90.4	99.6	99.6	99.4	99.4
0.3	50	12.6	10.8	13.2	12.4	13.0	23.1	25.4	27.7	26.1	26.4	37.2	46.7	48.6	45.6	45.7
	100	21.0	17.4	22.1	20.6	21.7	41.0	45.1	49.1	46.3	46.9	63.7	75.6	77.7	74.7	74.9
	150	29.3	23.9	30.8	28.7	30.3	56.4	61.4	66.0	63.0	63.7	80.7	90.1	91.6	89.7	89.9
	200	37.2	30.3	39.1	36.6	38.5	68.8	73.8	78.2	75.4	76.1	90.4	96.3	97.1	96.2	96.3
0.4	50	12.6	4.7	12.6	9.8	12.6	23.1	13.7	23.1	20.3	23.1	37.2	30.3	39.1	36.6	38.5
	100	21.0	6.0	21.0	16.3	21.0	41.0	23.1	41.0	36.5	41.0	63.7	53.2	66.3	63.1	65.5
	150	29.3	7.2	29.3	23.0	29.3	56.4	32.3	56.4	51.2	56.4	80.7	70.5	83.0	80.4	82.4
	200	37.2	8.3	37.2	29.9	37.2	68.8	41.0	68.8	63.7	68.8	90.4	82.3	92.0	90.3	91.6
0.6	50	12.6	0.6	12.6	8.7	12.6	23.1	2.5	23.1	17.2	23.1	37.2	8.3	37.2	29.9	37.2
	100	21.0	0.3	21.0	15.3	21.0	41.0	2.5	41.0	32.8	41.0	63.7	12.6	63.7	55.5	63.7
	150	29.3	0.2	29.3	22.2	29.3	56.4	2.5	56.4	47.7	56.4	80.7	16.9	80.7	74.3	80.7
	200	37.2	0.1	37.2	29.3	37.2	68.8	2.5	68.8	60.7	68.8	90.4	21.0	90.4	86.2	90.4

In the table, we omit the Δ_0 subscript for notational simplicity.

calculated using the subsample. This procedure is repeated 1,000 times. Among these repetitions, T_1 rejects the null hypothesis 71.5% of the time, i.e., the power of T_1 is 71.5%, while the combined test $T_{c,\Delta_0}(w)$ has power 82.4%, 74.2%, 71.1% when $\Delta_0 = 0.1, 0.2, 0.25$, respectively. Hence, the sensitivity parameter Δ_0 can be as large as 0.25 before the combined test starts to lose power compared to T_1. In comparison, $T_{2,\Delta_0}(w)$ has worse performance, with power equal to 80.4%, 64.8%, 45.7% when $\Delta_0 = 0.1, 0.2, 0.25$, respectively. Taking a closer look at the results, we note that if T_1 is larger than $c_{1-\alpha;\rho}$ defined in (4), then both T_1 and the combined test can reject H_0 regardless of the value of Δ_0. If $T_1 < z_{1-\alpha}$, then T_1 cannot reject H_0 while the combined test can still reject 27.7% of these cases at $\Delta_0 = 0.2$. The potential loss of using the combined test is when T_1 is between $z_{1-\alpha}$ and $c_{1-\alpha;\,\rho}$, in which cases using T_1 alone can reject H_0 but the combined test is sensitive to a certain value of Δ_0. However, this scenario is relatively rare and occurs in 8.4% of the repetitions; furthermore, even in this scenario, the combined test can still reject H_0 at $\Delta_0 = 0.2$ around half the time.

The last step of a sensitivity analysis is to reason about whether a value of $\Delta_0 = 0.2$ is plausible given that we have already controlled for baseline covariates listed in Table 1. For this task, an intuitive strategy is to judge the plausibility of Δ_0 in reference to some observed covariates [45]. Specifically, we can omit observed covariates one at a time during matching and calculate $\bar{Y}_0 - \bar{Y}_e$ using the resulting matched external controls. Using this procedure, we estimate the amount of bias from not matching on one of the observed covariates and to benchmark the plausibility of Δ_0, the amount of bias from not being able to match on the region variable. The results show that omitting the baseline HbA1c leads to the largest $\bar{Y}_0 - \bar{Y}_e$ that is equal to 0.14, while omitting any other observed variables in Table 1 leads to $\bar{Y}_0 - \bar{Y}_e$ ranging from −0.05 to 0.04. Based on the prior knowledge in the study of Home et al. [46] that the baseline HbA1c explains most of the variability in the change in HbA1c, particularly in comparison to the geographical region, we view that $\Delta = 0.2$ is implausible.

In summary, before looking at the data, the choice between T_1 and $T_{2,\Delta_0}(w)$, would be difficult to make or justify on the basis of a priori considerations. In some cases, T_1 may not be powerful enough due to the small sample size of the internal RCT, while leveraging external controls leads to a more powerful test. In some other cases, $T_{2,\Delta_0}(w)$ may be sensitive to unmeasured biases while T_1 is already powerful enough. Under these circumstances, the combined test $T_{c,\Delta_0}(w)$ is often preferable as it performs both tests with a small correction for multiple testing by taking into account the high correlation of the two test statistics.

6 Discussion

We propose a sensitivity analysis approach for using external controls in clinical trials to examine the robustness of study conclusion to remaining unmeasured bias after controlling for measured covariates. Results from the sensitivity analysis can be summarized by the “tipping point” – the magnitude of Δ_0 that would be needed such that the null hypothesis can no longer be rejected. If Δ_0 is deemed plausible (or implausible), the conclusion based on using external controls is sensitive (or robust) to unmeasured bias.

When in doubt about whether the use of external controls increases power, we propose a combined testing procedure that performs both tests, one only using the internal controls and one additionally using the external controls, correcting for multiple testing using the joint distribution of the two test statistics. Because the two test statistics are highly correlated, this correction for multiple testing is small, and thus the combined test only has a small loss of power compared to knowing a priori which test is best. Moreover, the combined test provides a new method of sensitivity analysis designed for data fusion problems, which anchors at the unbiased RCT-only analysis and spends a small proportion of the type I error to also test using the external controls. In this way, if leveraging external controls increases power, the power gain compared to the RCT-only analysis can be substantial; if not, the power loss is small.

Our work is motivated by the literature of sensitivity analysis, in which testing a hypothesis multiple times has been shown to be useful in enhancing the robustness to unmeasured bias [38,47–49]. Nonetheless, we focus on a distinct context and have shown that testing multiple times using both a known unbiased test and potentially biased tests can be particularly attractive for data fusion problems. We also have developed various properties of the combined procedure that has not appeared in the existing literature.

Finally, a remaining question is how to choose w for the combined test. The power of the combined test depends on w in a complicated way as w not only affects the definition of $T_{2,\Delta_0}(w)$ but also the correlation ρ, which makes finding the optimal w a cumbersome task. In practice, a reasonable choice is $w = \pi_0 n_r/(n_e + \pi_0 n_r)$, which minimizes the variance of $w\overline{Y}_0 + (1 - w)\overline{Y}_e$ when $\mathrm{Var}(Y^{(0)}|D = 1) = \mathrm{Var}(Y^{(0)}|D = 0)$. Another way is to pre-specify several values of w, calculate the corresponding test statistics, and combine all the test statistics using their joint null distribution. Because of the high correlation between these test statistics, the price paid for multiple testing will generally be small.

Acknowledgement: We would like to thank the two anonymous reviewers for helpful comments and suggestions.

Funding information: This work was supported by National Institute of Allergy and Infectious Diseases (NIAID 5 UM1 AI068617).

Conflict of interest: Y. Yi, Y. Zhang, and Y. Du are shareholders of Eli Lilly and Company.

References

[1] Jones DS, Podolsky SH. The history and fate of the gold standard. Lancet. 2015;385(9977):1502–3.

[2] Bothwell LE, Podolsky SH. The emergence of the randomized, controlled trial. N Engl J Med. 2016;375(6):501–4.

[3] Janes H, Donnell D, Gilbert PB, Brown ER, Nason M. Taking stock of the present and looking ahead: envisioning challenges in the design of future HIV prevention efficacy trials. Lancet HIV. 2019;6(7):e475–82.

[4] Sugarman J, Donnell DJ, Hanscom B, McCauley M, Grinsztejn B, Landovitz RJ. Ethical issues in establishing the efficacy and safety of long-acting injectable pre-exposure prophylaxis for HIV prevention: the HPTN 083 trial. Lancet HIV. 2021;8(11):e723–28.

[5] Rahman R, Ventz S, McDunn J, Louv B, Reyes-Rivera I, Polley M-YC, et al. Leveraging external data in the design and analysis of clinical trials in neuro-oncology. The Lancet Oncology. 2021;22(10):e456–65.

[6] Mintzer S, French JA, Perucca E, Cramer JA, Messenheimer JA, Blum DE, et al. Is a separate monotherapy indication warranted for antiepileptic drugs? The Lancet Neurology, 2015;14(12):1229–40.

[7] Eichler H-G, Pignatti F, Schwarzer-Daum B, Hidalgo-Simon A, Eichler I, Arlett P, et al. Randomized controlled trials versus real world evidence: neither magic nor myth. Clin Pharmacol Therapeutics. 2021;109(5):1212–8.

[8] Colnet B, Mayer I, Chen G, Dieng A, Li R, Varoquaux G, et al. Causal inference methods for combining randomized trials and observational studies: a review. 2020. arXiv: http://arXiv.org/abs/arXiv:2011.08047.

[9] Degtiar I, Rose S. A review of generalizability and transportability. 2021. arXiv: http://arXiv.org/abs/arXiv:2102.11904.

[10] Shi X, Pan Z, Miao W. Data integration in causal inference. 2021. arXiv: http://arXiv.org/abs/arXiv:2110.01106.

[11] Yang S, Zeng D, and Wang X. Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. 2020. arXiv: http://arXiv.org/abs/arXiv:2005.10579.

[12] Yang S, Zeng D, Wang X. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. 2020. arXiv: http://arXiv.org/abs/arXiv:2007.12922.

[13] Gagnon-Bartsch JA, Sales AC, Wu E, Botelho AF, Erickson JA, Miratrix LW, et al. Precise unbiased estimation in randomized experiments using auxiliary observational data. 2021. arXiv: http://arXiv.org/abs/arXiv:2105.03529.

[14] Chen S, Zhang B, Ye T. Minimax rates and adaptivity in combining experimental and observational data. 2021. arXiv: http://arXiv.org/abs/arXiv:2109.10522.

[15] Cheng D, Cai T. Adaptive combination of randomized and observational data. 2021. arXiv: http://arXiv.org/abs/arXiv:2111.15012.

[16] Li S, Luedtke A. Efficient estimation under data fusion. 2021. arXiv: http://arXiv.org/abs/arXiv:2111.14945.

[17] Li X, Miao W, Lu F, Zhou X-H. Improving efficiency of inference in clinical trials with external control data. 2020. arXiv: http://arXiv.org/abs/arXiv:2011.07234.

[18] Harton J, Segal B, Mamtani R, Mitra N, Hubbard R. Combining real-world and randomized control trial data using data-adaptive weighting via the on-trial score. 2021. arXiv: http://arXiv.org/abs/arXiv:2108.08756.

[19] Gao F, Glidden DV, Hughes JP, Donnell DJ. Sample size calculation for active-arm trial with counterfactual incidence based on recency assay. Stat Commun Infect Diseases. 2021;13(1):20200009.

[20] Liu Y, Lu B, Foster R, Zhang Y, John Zhong Z, Chen M-H, et al. Matching design for augmenting the control arm of a randomized controlled trial using real-world data. J Biopharmaceut Stat. 2022;32(1):1–17.

[21] Pocock SJ. The combination of randomized and historical controls in clinical trials. J Chronic Diseases. 1976;29(3):175–88.

[22] International Council for Harmonisation (ICH). Choice of control group and related issues in clinical trials E10. 2000.

[23] European Medicines Agency (EMA). Guideline on clinical trials in small populations. 2006.

[24] US Food and Drug Administration (FDA). Framework for FDA's real-world evidence program. 2018.

[25] Sharpless NE, Doroshow JH. Modernizing clinical trials for patients with cancer. JAMA Feb 2019;321 (5):447–8.

[26] Carrigan G, Whipple S, Capra WB, Taylor MD, Brown JS, Lu M, et al. Using electronic health records to derive control arms for early phase single-arm lung cancer trials: proof-of-concept in randomized controlled trials. Clin Pharmacol Therapeut. 2020;107(2):369–77.

[27] Schmidli H, Häring DA, Thomas M, Cassidy A, Weber S, Bretz F. Beyond randomized clinical trials: use of external controls. Clin Pharmacol Therapeutics. 2020;107(4):806–16.

[28] Thorlund K, Dron L, Park JJH, Mills EJ. Synthetic and external controls in clinical trials-a primer for researchers. Clin Epidemiol. 2020;12:457–67.

[29] Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. J R Stat Soc. Ser A (Stat Soc). April 2011;174(2):369–86.

[30] Chen M-H, Ibrahim JG. Power prior distributions for regression models. Stat Sci. 2000;15(1):46–60.

[31] Nikolakopoulos S, van der Tweel I, Roes KCB. Dynamic borrowing through empirical power priors that control type i error. Biometrics. 2018;74(3):874–80.

[32] International Council for Harmonisation (ICH). Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. 2019.

[33] Hirose T, Cai Z, Yeo KP, Imori M, Ohwaki K, Imaoka T. Open-label, randomized study comparing basal insulin peglispro and insulin glargine, in combination with oral antihyperglycemic medications, in Insulin-Naïve Asian patients with type 2 diabetes. J Diabetes Investigat. 2018;9(1):100–7.

[34] Davies MJ, Russell-Jones D, Selam J-L, Bailey TS, Kerényi Z, Luo J, et al. Basal insulin peglispro versus insulin glargine in Insulin-Naïve type 2 diabetes: imagine 2 randomized trial. Diabetes Obesity Metabolism. 2016;18(11):1055–64.

[35] Lim JU, Lee JH, Kim JS, Hwang YI, Kim T-H, Lim SY, et al. Comparison of world health organization and Asia-Pacific body mass index classifications in COPD patients. Int J Chronic Obstruct Pulmonary Disease. 2017;12:2465–75.

[36] Rosenbaum PR. Design of observational studies, (2nd edn.), New York, NY: Springer; 2020.

[37] Rosenbaum PR. Observational studies. New York, NY: Springer; 2002.

[38] Rosenbaum PR. Testing one hypothesis twice in observational studies. Biometrika. 2012;99(4):763–74.

[39] Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. J Amer Stat Assoc. 1980;75(371):591–3.

[40] Cox DR, Spjøtvoll E, Johansen S, van Zwet WR, Bithell JF, Barndorff-Nielsen O, et al. The role of significance tests [with discussion and reply]. Scandinavian J Stat. 1977;4(2):49–70.

[41] Austin PC, Small DS. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. Stat Med. 2014;33(24):4306–19.

[42] Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods. 2008;13(4):279–313.

[43] Rosenbaum PR. Design sensitivity in observational studies. Biometrika. 2004;91(1):153–64.

[44] Cornfield J, Haenszel W, Hammond E, Lilienfeld A, Shimkin M, Wynder E. Smoking and lung cancer. J Nat Cancer Institute. 1959;22:173–203.

[45] Imbens GW. Sensitivity to exogeneity assumptions in program evaluation. Amer Econ Rev Papers Proc. May 2003;93(2):126–32.

[46] Home PD, Shen C, Hasan MI, Latif ZA, Chen JW, González Gálvez G. Predictive and explanatory factors of change in hba1c in a 24-week observational study of 66,726 people with type 2 diabetes starting insulin analogs. Diabetes Care. 2014;37(5):1237–45.

[47] Small DS, Cheng J, Halloran ME, Rosenbaum PR. Case definition and design sensitivity. J Amer Stat Assoc. Jan 2013;108(504):1457–68.

[48] Rosenbaum PR, Small DS. An adaptive mantel-haenszel test for sensitivity analysis in observational studies. Biometrics. 2017;73(2):422–30, 2021/03/16.

[49] Ye T, Small DS. Combining broad and narrow case definitions in matched case-control studies. 2021. arXiv: http://arXiv.org/abs/arXiv:2105.01124.