Estimation of causal effects in clinical endpoint bioequivalence studies in the presence of intercurrent events: noncompliance and missing data

Yiyue Lou, Michael P. Jones & Wanjie Sun

Taylor & Francis
Taylor & Francis Group

Check for updates

Estimation of causal effects in clinical endpoint bioequivalence studies in the presence of intercurrent events: noncompliance and missing data

Yiyue Lou[a], Michael P. Jones[a], and Wanjie Sun[b]

[a]Department of Biostatistics, University of Iowa College of Public Health, Iowa City, IA, USA; [b]Office of Biostatistics, Center for Drug Evaluation and Research, Food and Drug Administration (CDER/FDA), Silver Spring, MD, USA

ABSTRACT

In clinical endpoint bioequivalence (BE) studies, the primary analysis for assessing equivalence between a generic and an innovator product is based on the observed per-protocol (PP) population (usually completers and compliers). However, missing data and noncompliance are post-randomization intercurrent events and may introduce selection bias. Therefore, PP analysis is generally not causal. The FDA Missing Data Working Group recommended using "causal estimands of primary interest." In this paper, we propose a principal stratification causal framework and co-primary causal estimands to test equivalence, which was also recommended by the recently published ICH E9 (R1) addendum to address intercurrent events. We identify three conditions under which the current PP estimator is unbiased for one of the proposed co-primary causal estimands – the "Survivor Average Causal Effect" (SACE) estimand. Simulation shows that when these three conditions are not met, the PP estimator is biased and may inflate Type 1 error and/or change power. We also propose a tipping point sensitivity analysis to evaluate the robustness of the current PP estimator in testing equivalence when the sensitivity parameters deviate from the three identified conditions, but stay within a clinically meaningful range. Our work is the first causal equivalence assessment in equivalence studies with intercurrent events.

Introduction

In clinical endpoint bioequivalence (BE) studies evaluating the equivalence between a generic drug and an innovator drug, intercurrent events such as missing data and noncompliance are commonly observed. The current primary equivalence analysis is usually based on the PP population (i.e., completers and compliers in general) to evaluate the "net" treatment effect. For superiority trials, the intent-to-treat (ITT) analysis, which utilizes all subjects regardless of compliance status, is usually preferred because it is conservative in rejecting the null hypothesis of no difference. For equivalence or non-inferiority trials, however, as Gupta (2011) stated, "It has been argued that protocol violations and poorly conducted trials may cause the results obtained from two different treatment groups to appear similar. ITT tends to make the two treatments look similar, whereas the PP remove patients who do not complete treatment and is more able to reflect treatment differences." Hence, ITT analysis is generally anti-conservative for equivalence or noninferiority trials (Sanchez et al 2006; Snapinn, 2000), and the PP population is usually preferred in equivalence assessment. As the previous FDA noninferiority guidance discussed, "Medication non-compliance or misclassification/measurement error, errors that would be fatal to

success in a superiority study, can lead to apparently favourable (results) in a non-inferiority study" (FDA, 2010). However, Frangakis and Rubin (2002) pointed out that treatment noncompliance and missingness are posttreatment (post-randomization), and hence may be consequences of treatment. Therefore, PP analysis is subject to selection bias, and a crude comparison within the observed PP population between the two treatment groups is generally not a causal effect.

In the recently published ICH E9 Revision 1 (R1) addendum (ICH, 2017; Aug 30, 2017), the importance of intercurrent events was highlighted. The ICH E9 (R1) working group warned that "choosing and defining efficacy and safety variables as well as standards for data collection and methods for statistical analysis without first addressing the occurrence of intercurrent events will lead to ambiguity about the treatment effect to be estimated and potentially misalignment with trial objectives." Intercurrent events are "events that occur after treatment initiation and either preclude observation of the variable or affect its interpretation," such as discontinuation of treatment, use of an alternative treatment, and others. The ICH E9 (R1) addendum stated that "estimation of a treatment effect from any analysis where membership is based on intercurrent events on the assigned treatments is liable to confounding because different subjects will experience different intercurrent events on different treatments." The addendum revisited "the meaning and role of the per-protocol (PP) analysis, in particular whether the need to explore the impact of protocol violations and deviations can be addressed in a way that is less biased and more interpretable than naïve analysis of the per protocol set."

The ICH E9 (R1) addendum also emphasized the importance of defining estimand and using sensitivity analysis, which were previously discussed by the Missing Data Working Group of the U.S. Food and Drug Administration (FDA) (LaVange and Permutt, 2016; Permutt, 2016a, 2016b) as well. The FDA Missing Data Working Group also recommended that the trial protocol should explicitly define the measures of intervention effects, i.e., the causal estimands of primary interest.

For equivalence trials, this poses a dilemma: on the one hand, "net" treatment effect among the PP population (i.e., completers and compliers) is preferred as a conservative approach to assess equivalence; on the other hand, PP is an intercurrent event that may introduce selection bias. Resolving this dilemma is a pressing task for the analysis of clinical endpoint BE studies.

In this article, we aim to address two issues: 1) how to define causal estimands for equivalence assessment in clinical endpoint BE studies in the presence of intercurrent events – noncompliance and missing data, and 2) what primary and sensitivity analyses approaches can be used to estimate the proposed causal estimands for equivalence assessment. We have organized this article as follows. In the Methods section, we propose a causal framework by applying Frangakis and Rubin's principal stratification; co-primary causal estimands for assessing equivalence in clinical endpoint BE studies; and primary analysis and a tipping point sensitivity analysis for equivalence assessment. We also quantify the bias of the current PP estimand from the proposed principal stratum casual estimand SACE and evaluate conditions under which the two are two equal, i.e., bias is zero. In the Simulation section, we demonstrate the impact on Type 1 error and power under various scenarios. In the Case Study section, we apply the proposed causal estimands and primary and sensitivity analyses to a clinical endpoint BE study for acne vulgaris registered at ClinicalTrials.gov. Finally, we conclude with some discussion of the contributions of this paper and future research areas.

Methods

Clinical endpoint BE study

For systemic generic drugs, pharmacokinetic (PK) studies are usually used to assess BE, i.e., the rate and extent of which the active ingredient or active moiety in pharmaceutical equivalents or pharmaceutical alternatives becomes available at the site of drug action, between a generic drug and an innovator drug. For locally acting drugs where systemic exposure may not necessarily be relevant, clinical endpoint BE studies are usually used to determine BE (Grosser et al., 2015; Lawrence and Li, 2014).

In a clinical endpoint BE study, one or several predetermined clinical endpoints are used to evaluate comparative clinical effects between a generic drug (TEST) and an innovator drug or reference-listed drug (RLD) in the chosen population. The general design of these studies is a double-blind, randomized, parallel study with multiple clinical visits. A placebo (PLB) or vehicle arm is usually included in these studies in order to demonstrate that the study is sufficiently sensitive to identify the clinical effects in the enrolled patient population (Lawrence and Li, 2014).

After randomization, participants may drop out of a clinical endpoint BE study due to lack of efficacy, adverse events, lost to follow-up, or for non-study-related reasons. Another complication is treatment noncompliance, such as poor/hyper-treatment compliance, missing doses, out-of-window visits, or taking restricted medications. According to a previous meta-analysis using clinical endpoint BE studies for topical drugs, dropouts and noncompliance did not occur completely at random, but were related to the treatment effect, and were positively correlated, i.e., those dropped out were more likely to be noncompliant and vice versa (Sun et al., 2016).

To establish BE between TEST and RLD, a study has to pass two superiority tests (TEST vs. PLB and RLD vs. PLB) for validation of the sensitivity of the study and one equivalence test (TEST vs. RLD) for establishment of average equivalence. Let μ_T and μ_R denote the mean outcomes under TEST and RLD, where the outcomes can be continuous or binary or of other types. The equivalence analysis usually compares either the ratio of means or the difference of means (on the original scale or after log transformation) using two one-sided tests (Schuirmann, 1987; FDA, 2001; Sun et al., 2017). In this paper, we focus on the difference of means as the measure of treatment effect. The compound hypothesis to be tested is:

$$H_0: \mu_T - \mu_R \leq \theta_1 \text{ or } \mu_T - \mu_R \geq \theta_2 \text{vs. } H_a: \theta_1 < \mu_T - \mu_R < \theta_2,$$

where $[\theta_1, \theta_2]$ are the prespecified equivalence margins. We reject the null hypothesis if the 90% confidence interval (CI) for the difference of means is contained within $[\theta_1, \theta_2]$.

Causal inference and notation

Let A denote the treatment assignment (1 = TEST, 0 = RLD) and Y denote the observed outcome for a subject. For simplicity, subscripts for subjects are omitted here. According to the Rubin Causal Model, every subject could potentially receive both TEST and RLD. Let Y_1 and Y_0 denote the potential outcomes for a subject at a particular point in time had that subject received TEST or RLD, respectively. Y_1 and Y_0 are considered "pretreatment" or "pre-randomization" since they exist before the actual treatment assignment (A). In other words, randomization or treatment assignment does not affect the potential outcomes. However, in parallel studies, after treatment assignment, we can only observe one potential outcome for each subject depending on the treatment group to which the subject is assigned. A causal effect of treatment assignment A on outcome Y is defined to be a comparison between the potential outcomes Y_1 and Y_0 on a common set of units (Neyman, 1923; Rubin, 1974, 1978).

The population average causal effect of A on Y is defined as a comparison between the average potential outcomes Y_1 and Y_0 in the entire population, i.e., $E(Y_1) - E(Y_0)$. In a completely randomized clinical trial, if there is no intercurrent event, the population average causal effect is equivalent to the ITT estimand, i.e.,

$$E(Y_1) - E(Y_0) = E(Y| A = 1) - E(Y|A = 0). \tag{1}$$

This is because under random assignment, subjects assigned to each treatment group are considered identical or exchangeable (Rubin, 1977) in the sense that the average potential outcome under TEST is the same for subjects assigned to TEST or RLD, i.e., $E(Y_1|A = 1) = E(Y_1|A = 0) = E(Y_1)$. According to the consistency assumption (Robins et al., 2000), if we believe that the observed outcome when a subject is assigned to TEST, $Y|A$, is precisely his/her potential outcome under TEST, $Y_1|A$, the mean observed outcome in the group assigned to TEST equals the mean potential outcome under TEST, i.e., $E(Y|A = 1) = E(Y_1)$. The RLD arm is handled similarly.

We now let S denote the observed PP status ($s =$ complier and completer, i.e., PP; $\bar{s} =$ non-complier or dropout, i.e., non-PP) for a subject, and let S_1 and S_0 denote the potential PP status had a subject been assigned to TEST or RLD, respectively. Since the observed PP status (S) occurs after randomization or treatment assignment A and before the observed outcome Y, it is an intermediate or intercurrent variable, which is "post-randomization."

The current estimand for equivalence analysis, which we call the "observed PP estimand," is the difference of the mean observed outcome Y among the observed PP population who are assigned to TEST and RLD:

$$\delta_{PP} = E(Y|A = 1, \ S = s) - \ E(Y|A = 0, \ S = s). \tag{2}$$

δ_{PP} is also called the "net treatment effect" of treatment assignment A adjusting for the posttreatment variable S (i.e., the observed PP status) (Cochran, 1957; Rosenbaum, 1984). If A is completely randomized, the observed PP estimand is equal to

$$\delta_{PP} = E(Y_1|S_1 = s) - E(Y_0|S_0 = s) \tag{3}$$

which is the difference of the mean potential outcomes under TEST and RLD among those who would be PP if assigned to TEST ($S_1 = s$) and those who would be PP if assigned to RLD ($S_0 = s$). This net treatment effect is problematic because the potential PP under the TEST group ($S_1 = s$) is not necessarily the same group of subjects as those potential PP under RLD ($S_0 = s$), i.e., it may not be a common set of units (Neyman, 1923; Rubin, 1974, 1978). For example, a subject who can tolerate RLD may not be able to tolerate TEST and may drop out of the study. Therefore, the current estimand for the equivalence analysis (δ_{PP}) is usually not a causal estimand.

Proposed causal principal strata framework for equivalence assessment in clinical endpoint BE studies

To evaluate causal effects in the presence of posttreatment variables, Frangakis and Rubin (2002) proposed a general framework of principal stratification, which had been previously introduced by Angrist et al. (1996) and Robins (1986). "Principal strata" are defined by the joint potential values of the posttreatment intermediate variable under each treatment. The key property of principal strata is that they are not affected by treatment assignment; hence, they can be used as pre-randomization variables such as age or gender. Because of this property, principal effects, i.e., the treatment effects within principal strata, are always causal effects. Principal stratification in causal inference has been used to address a broad class of problems concerning "censoring by death" (e.g., QOL is undefined for those who died before the outcome is recorded), treatment noncompliance, and surrogate outcomes (Frangakis and Rubin, 2002). Our approach was also recommended by the recently published ICH E9 (R1) addendum (International Conference on Harmonisation, 2017), which included principal stratification as one of the strategies to address intercurrent events.

We propose a causal framework for equivalence assessment in clinical endpoint BE studies by applying Frangakis and Rubin's (2002) principal stratification. Let U denote the principal stratum defined by the joint potential PP status had a subject been assigned to TEST or RLD, i.e., $U = \{S_1, \ S_0\} = \{ss, \ s\bar{s}, \ \bar{s}s, \ \bar{s}\bar{s}\}$. There are four principal strata.

- "Always PP" ($U = ss$): Participant who would comply with and complete the study (i.e., be PP) under both TEST and RLD.
- "PP with TEST only" ($U = s\bar{s}$): Participant who would be PP if assigned to TEST, but would not if assigned to RLD.
- "PP with RLD only" ($U = \bar{s}s$): Participant who would be PP if assigned to RLD, but would not if assigned to TEST.
- "Never PP" ($U = \bar{s}\bar{s}$): Participant who would not be PP regardless of the treatment group being assigned to.

This principal stratum framework changes the original post-randomization intercurrent variable (i.e., the observed PP status) between treatment assignment (A) and observed outcome (Y) to a pre-randomization variable (i.e., the potential PP status), which is not affected by treatment assignment, similar to age and gender (Frangakis and Rubin, 2002). For example, whether or not one subject can tolerate TEST has no relationship to the treatment group s/he is assigned to. Therefore, we can now evaluate causal treatment effect on the potential outcome within the same pre-randomization principal stratum, $E(Y_1 - Y_0 | U = u, u = (ss, \bar{s}s, s\bar{s}, \bar{s}\bar{s}))$, i.e., a principal causal effect, which is always causal according to Frangakis and Rubin (2002).

Figure 1 is a graphical representation of treatment assignment (A), observed intermediate, or intercurrent variable – PP status (S), and the proposed principal strata (U) for clinical endpoint BE studies. For example, if a subject is assigned to TEST (A = 1) and complies with and completes the study (S = s), the potential outcome under TEST (Y_1) is observed (but Y_0 is not). This subject can either be an always PP ($U = ss$) or be a PP with TEST only ($U = s\bar{s}$), which can also be seen in Figure 2. In the latter case, s/he may not tolerate RLD due to adverse effects or may drop out of the study due to lack of efficacy or for other treatment-related reasons. However, based on the observed data, we cannot distinguish between always PP ($U = ss$) and PP with TEST ($U = s\bar{s}$), only among those observed PP (S = s) in TEST. Likewise, if a subject is assigned to TEST but does not comply with or complete the study ($S = \bar{s}$), this subject can belong to either PP with RLD only or never PP. In contrast to PP with TEST only or PP with RLD only, which are always related to treatment, never PP can occur due to either study-related reasons or non-study-related reasons (e.g., a subject moves out of the town).
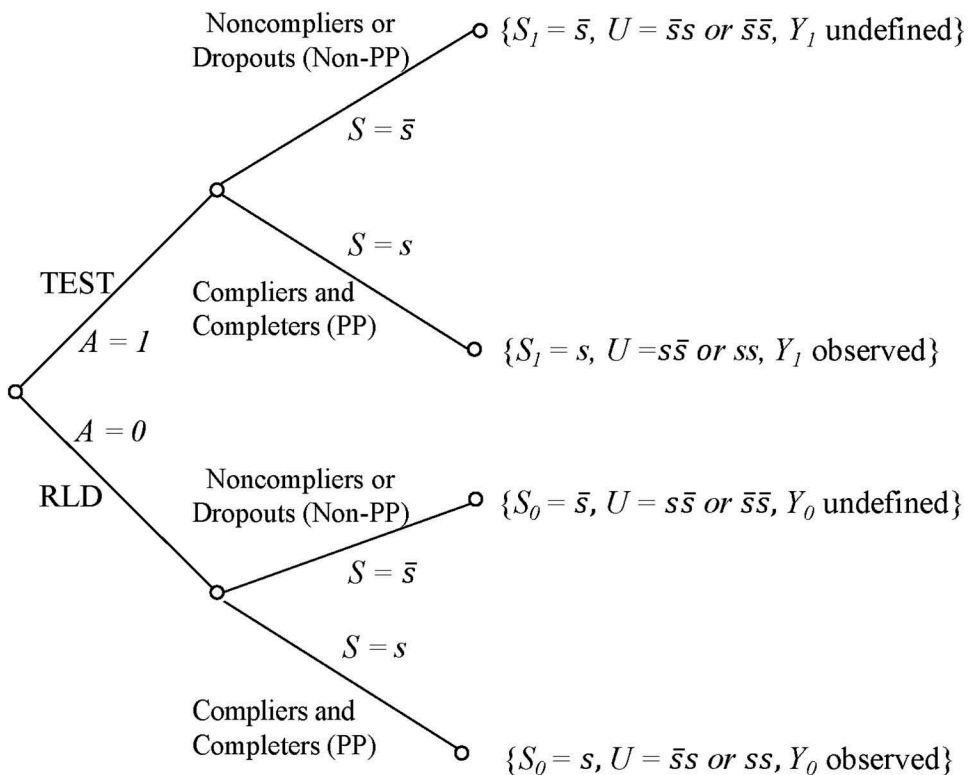

Figure 1. Treatment assignment (A), observed PP status (S), and principal strata (U) in clinical endpoint BE studies with missing and noncompliance data.
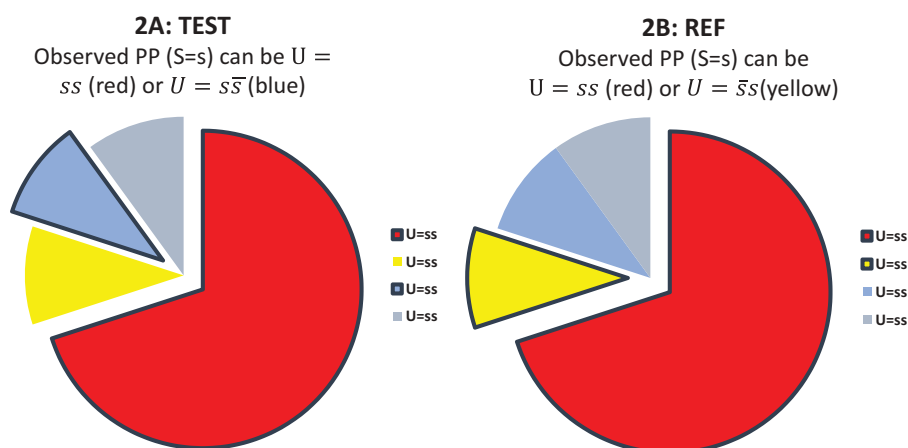
Figure 2. Observed PP status (bordered) in TEST (2a) and RLD (2b) stratified by four principal strata based on potential PP status.

Proposed co-primary causal estimands for equivalence assessment in clinical endpoint BE studies

Within the framework of principal stratification causal inference, a meaningful effect for evaluating equivalence is the causal treatment effect comparison among those who would be PP under both TEST and RLD, i.e., within the principal stratum of always PP: $E(Y_1|U = ss) - E(Y_0|U = ss)$. Rubin (2006) called this "the survivor average causal effect" (SACE) in the example of "censoring by death," i.e., the treatment effect among the always survivor group. SACE is a meaningful comparison of treatment effects for assessing equivalence for two reasons.

For one, SACE measures the desired "net treatment effect:" equivalence between TEST and RLD among the always PP stratum, which is in line with the current observed PP estimand for equivalence testing. However, SACE is an improvement over the observed PP estimand in that it is a causal effect: SACE compares the treatment effect on a common set of units – the principal stratum of always PP ($U = ss$, red pies in Figure 2). On the contrary, the observed PP estimand compares treatment effect not necessarily on a common set of units. As shown in Figure 2, the TEST observed PP is a union of two principal strata: always PP ($U = ss$, red in Figure 2(a)) and TEST PP only ($U = s\bar{s}$, blue in Figure 2(a)), whereas the RLD-observed PP is a union of two principal stratum: always PP ($U = ss$, red in Figure 2(b)) and RLD PP only ($U = \bar{s}s$, yellow in Figure 2(b)). Second, SACE excludes the unobserved/missing outcomes or nonexistent outcomes where dropout/ noncompliance reasons may or may not be fully observed in practice, which avoids imputation or ambiguity and simplifies defining the causal estimand for equivalence assessment.

Furthermore, as Sun et al. (2016) reported, oftentimes missing data and noncompliance data are not completely at random, but related to treatment effect. This means that being non-PP or PP itself is an outcome. Hence, to conclude that the two drugs are equivalent, besides evaluating SACE (the net treatment effect within the principal stratum of always PP), we can also evaluate equivalence between the two treatment groups in the proportion of being non-PP or PP, which captures the remaining outcome information and is a simple and indirect coverage of the other three principal strata.

Therefore, we propose co-primary causal estimands for the equivalence analysis in clinical end-point BE studies as follows.

(1) The SACE estimand: The difference in the average potential outcome between TEST and RLD within the principal stratum of always PP:

$$\delta_{SACE} = E(Y_1|U = ss) - E(Y_0|U = ss). \tag{4}$$

The null and alternative hypotheses for equivalence based on the SACE estimand are, respectively, as follows:

$$H_0: \delta_{SACE} \leq \theta_1 \text{ or } \delta_{SACE} \geq \theta_2;$$

$$H_1: \theta_1 < \delta_{SACE} < \theta_2,$$

where θ_1 and θ_2 are prespecified equivalence margins.

(2) The proportion estimand: The difference in the proportion of being potential non-PP between TEST and RLD:

$$\delta_p = Pr(S_1 = \bar{s}) - Pr(S_0 = \bar{s}). \tag{5}$$

In a completely randomized clinical trial, similar to the ITT estimand, the proportion estimand (δ_p) can be directly estimated from the observed data, i.e., the difference in the proportion of observed PP between TEST and RLD:

$$\delta_p = Pr(S = \bar{s}|A = 1) - Pr(S = \bar{s}|A = 0). \tag{6}$$

Likewise, we can specify the hypotheses for equivalence assessment based on the proportion estimand:

$$H_0: \delta_P \leq \theta_1 \text{ or } \delta_P \geq \theta_2;$$

$$H_1: \theta_1 < \delta_P < \theta_2.$$

Proposed primary and sensitivity analyses methods for SACE estimand

Estimating the SACE estimand is not straightforward, because the observed PP population under TEST (or RLD) is a mixture of always PP and PP with TEST (or RLD) only (Figure 2). Without additional assumptions, we cannot identify who is always PP and who is PP with TEST (or RLD) only. Current methods for assessing SACE and other principal strata causal effects fall into three categories: 1) imposing additional assumptions (e.g., ignorability or monotonicity) for estimation of SACE (Ding et al., 2011; Frumento et al., 2012; Zhang et al., 2009); 2) deriving bounds for principal strata causal effect (Chiba et al., 2012; Imai, 2008; Rubin, 2006; Yang and Small, 2016; Zhang and Rubin, 2003); and 3) using sensitivity analysis techniques (Chiba and VanderWeele, 2011; Egleston et al.; 2007; Hayden et al., 2005).

For clinical endpoint BE studies, rather than imposing untestable assumptions (e.g., ignorability or missing at random) that are hard to justify in clinical trials or non-applicable assumptions for equivalence trials (e.g., monotonicity), we propose to use Chiba and VanderWeele's (2011) sensitivity approach, which relies on minimal assumptions that can be easily satisfied by randomized clinical trials.

Assumption 1: The stable unit treatment value assumption (SUTVA)
A subject's potential outcomes under treatment remain the same regardless of the mechanism used to assign the subject to treatment and regardless of the treatment assignments of other subjects (Rubin, 1980).

SUTVA contains two parts: 1) only one single version of each treatment, i.e., there are no hidden variations in treatment, and the potential outcomes are well defined for each treatment; and 2) no

interference between units (Cox, 1958), and thus the potential outcomes of any subject are assumed to be unaffected by the treatment assignment of any other subject.

Assumption 2: Random assignment of treatment

The treatment assignment is completely independent of potential outcomes,

$A \coprod^{S_0}, S_1, Y_0, Y_1,$ *or equivalently,* $Pr(A|S_0, S_1, Y_0, Y_1) = Pr(A)$ *is constant across subjects.*

Under the assumptions of SUTVA and random assignment, Chiba and VanderWeele (2011) demonstrated that the SACE estimand (4) is composed of two parts:

$$
\begin{aligned}
\delta_{SACE} =\ & E(Y_1|\ U = ss) - E(Y_0|\ U = ss) \\
=\ & E(Y|\ A = 1,\ S = s) - E(Y|A = 0,\ S = s) + \frac{\pi_{\bar{s}s}}{p_0}\beta_0 - \frac{p_1 - p_0 + \pi_{\bar{s}s}}{p_1}\beta_1 \\
=\ & \delta_{PP} + Bias.
\end{aligned} \tag{7}
$$

In the first part of equation (1), $E(Y|\ A = 1,\ S = s) - E(Y|A = 0,\ S = s)$ is simply the observed PP estimand, δ_{PP} in (2), which can be unbiasedly estimated from the observed data. The second part is the bias of the observed PP estimand, δ_{PP}, deviating from the true causal effect of primary interest, SACE δ_{SACE} (4). The bias component involves three unknown but fixed sensitivity parameters: $\beta_0 = E(Y_0|U = \bar{s}s) - E(Y_0|U = ss)$ is the difference in average potential outcome under RLD between the stratum "PP with RLD only" and the stratum "always PP;" $\beta_1 = E(Y_1|U = s\bar{s}) - E(Y_1|U = ss)$ is the difference in average potential outcomes under TEST between the stratum "PP with TEST only" and the stratum "always PP;" and $\pi_{\bar{s}s} = Pr(U = \bar{s})$ is the marginal proportion of PP with RLD only. Besides these three sensitivity parameters, $p_1 = Pr(S = s|\ A = 1)$ and $p_0 = Pr(S = s|\ A = 0)$ are the proportions of being observed PP under TEST and RLD, respectively, which can be directly estimated from the observed data.

We propose to continue using the estimator for the observed PP estimand, δ_{PP} (i.e., the PP estimator), as the primary analysis to estimate the causal SACE estimand, δ_{SACE}, as in current practice. However, we also propose to use a tipping point sensitivity analysis to evaluate the robustness of the conclusion based on the primary analysis (observed PP estimator) under different scenarios, by varying the values of β_0, β_1, and $\pi_{\bar{s}s}$ within clinically meaningful ranges.

The boundaries of $\pi_{\bar{s}s}$ can be easily derived as

$$
\pi_{\bar{s}s} \in [max(0,\ p_0 - p_1),\ min(p_0,\ 1 - p_1)] \tag{8}
$$

The proportion of PP with RLD only ($\pi_{\bar{s}s}$) is bounded, but the selection effects (β_0, β_1) are not. Furthermore, because the sensitivity parameters are considered as "fixed," the 90% CI for an estimator of the SACE can simply be constructed by shifting the CI of the observed PP estimand by the bias $\frac{\pi_{\bar{s}s}}{p_0}\beta_0 - \frac{p_1 - p_0 + \pi_{\bar{s}s}}{p_1}$ as in (7). Therefore, different equivalence conclusions can be drawn at different values of the sensitivity parameters. In particular, we are interested in the tipping point, where the equivalence conclusion changes directions, e.g., from passing BE to failing BE. Clinicians can then evaluate the plausibility of the corresponding values of sensitivity parameters β_0, β_1, and $\pi_{\bar{s}s}$. If plausible, we need to reconsider the study conclusion from the primary analysis, even though BE may have been established based on the observed PP estimator.

When is the Observed PP Estimand (δ_{PP}) Equal to the SACE Estimand (δ_{SACE})?

Is there any situation in which the bias in (7) is zero, i.e., the observed PP estimand for mean difference, $\delta_{PP} = E(Y|A = 1, S = s) - E(Y|A = 0, S = s)$ (2), is equal to the SACE estimand δ_{SACE} (4)? The answer is yes. We identified the following three conditions in which $\delta_{SACE} = \delta_{PP}$.

Condition 1: Direct Effects Only ($\pi_{s\bar{s}} = \pi_{\bar{s}s} = 0$)

δ_{SACE} equals δ_{PP} if $\pi_{s\bar{s}} = \pi_{\bar{s}s} = 0$. Note that $\pi_{\bar{s}s} + \pi_{ss} = p_0$ and $\pi_{s\bar{s}} + \pi_{ss} = p_1$, so that $\pi_{s\bar{s}} = p_1 - p_0 + \pi_{\bar{s}s}$. This means that all subjects are either always PP or never PP under TEST and RLD, and no subjects are TEST only or RLD only. This suggests a strong assumption that being PP or not is random and not related to the treatment assignment. Therefore, treatment effect on the outcome is "direct" and does not operate through the intermediate or intercurrent observed PP status (VanderWeele, 2015). However, Sun et al.'s (2016) meta-analysis showed that dropout and noncompliance are significantly associated with the treatment effect. Therefore, "Direct Effects Only" may not be a realistic condition.

Condition 2: No Selection Effects ($\beta_0 = \beta_1 = 0$)

δ_{SACE} equals δ_{PP} if $\beta_0 = \beta_1 = 0$. If $E(Y_0|U = \bar{s}s) = E(Y_0|U = ss)$ (i.e., $\beta_0 = 0$), the average potential outcome under RLD is the same regardless of whether a subject is PP with RLD only or always PP. Likewise, if $\beta_1 = E(Y_1|U = s\bar{s}) - E(Y_1|U = ss) = 0$, the average potential outcome under TEST is the same regardless of whether a subject is PP with TEST only or always PP. However, this assumption is strong and generally considered as unacceptable, because the principal strata subpopulations may differ on various characteristics related to the outcome (Little et al., 2009). For example, for acne drugs, a subject who would tolerate only the RLD may have a different skin type compared to a subject who would tolerate both the RLD and the generic drug TEST, and therefore may have a different treatment effect and different potential outcomes under RLD. White (2005) called β_0 and β_1 "selection effects." Little et al. (2009) pointed out that this assumption can be weakened by adjusting for an adequate set of known covariates (see also Chiba and VanderWeele (2011) and Ding and Lu (2017)). However, in clinical trials, it is generally hard to know under which condition it is considered "adequate."

Condition 3: Ideal-Exact Equality ($\beta_0 = \beta_1$ and $\pi_{s\bar{s}} = \pi_{\bar{s}s}$)

If $\beta_0 = \beta_1$ and $\pi_{s\bar{s}} = \pi_{\bar{s}s}$, the two parts of the sensitivity component of (7) cancel each other out, and SACE δ_{SACE} equals the first part – the observed PP estimand, δ_{PP}. What this means to clinical endpoint BE studies is that, if the proportion of observed PP is the same between TEST ($p_1 = Pr(S = s| A = 1) = \pi_{s\bar{s}} + \pi_{ss}$) and RLD ($p_0 = Pr(S = s| A = 0) = \pi_{\bar{s}s} + \pi_{ss}$), and the selection effects on TEST (β_1) and RLD (β_0) are the same, the estimator for the observed PP estimand is also unbiased for the SACE estimand of interest. This happens if TEST and RLD have exactly the same efficacy and safety, which is an "ideal-exact equality" scenario.

Intuitively, in Figure 3, these three conditions either block the first pathway from treatment assignment (A) to the intercurrent variable – observed PP (S) by direct effect only (Condition 1) – or block the second pathway from the intercurrent variable to the observed outcome (Y) by no selection effect (Condition 2) or create identical pathways from A to S and from S and Y by ideal exact equality (Condition 3). The underlying assumption of the current (and proposed) primary analysis (i.e., using the observed PP estimator to estimate the SACE estimand) is that at least one of these conditions is satisfied. However, in real-life clinical trials, none of these three conditions may easily be satisfied.

Simulation

Simulation is used to examine the finite sample performance of different estimators for the SACE estimand, to empirically validate the three identified conditions for the observed PP estimand (δ_{PP}) to be equal to the SACE estimand (δ_{SACE}), and to evaluate the bias of the PP estimator from the true SACE estimand and the impact on Type 1 error and power when these three conditions are not satisfied.
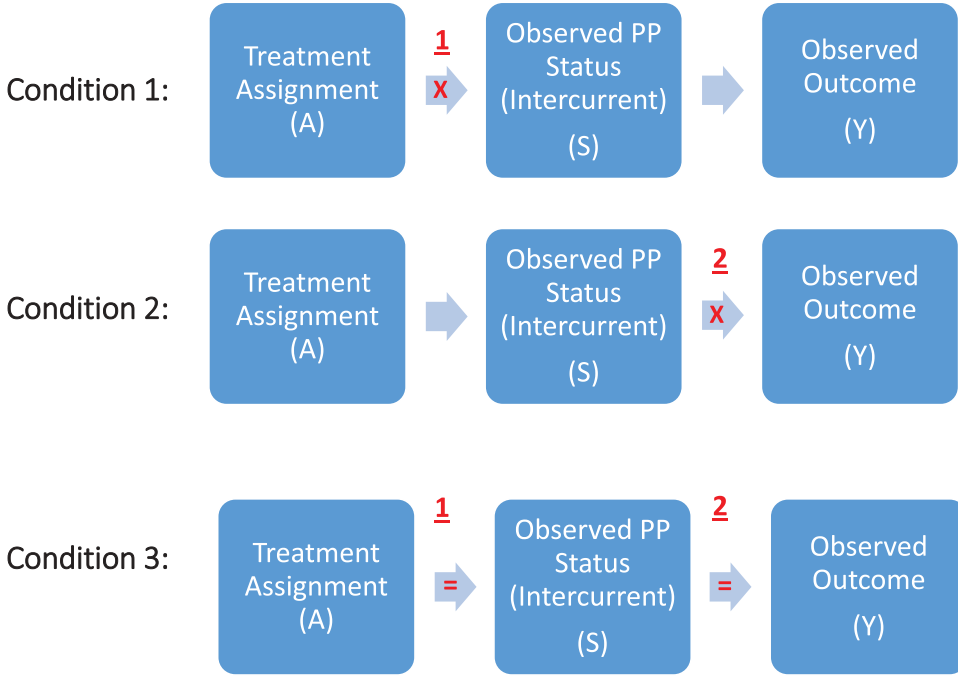
$$\delta_{SACE} = \delta_{PP}$$

Figure 3. Three conditions when $\delta_{SACE} = \delta_{PP}$.

Data generation

We generate data under the principal stratification framework. In particular, normal and binary potential outcome data are generated by the following steps.

- Generate principal strata U: For the ith subject, we generate principal strata membership U_i based on an independent multinomial distribution with probabilities $(\pi_{ss}, \ \pi_{s\bar{s}}, \pi_{\bar{s}s}\pi_{\bar{s}\bar{s}})$.
- Generate random treatment assignment A: $A_i \sim Bernoulli(1/2)$.
- Generate observed PP status S: If $A_i = 1$ and $U_i = ss$ or $U_i = s\bar{s}$, then $S_i = s$; similarly, if $A_i = 0$ and $U_i = ss$ or $U_i = \bar{s}s$, then $S_i = s$. Otherwise, $S_i = \bar{s}$.
- Generate potential outcomes Y_1 and Y_0: Here, only those potential outcomes that are used in Chiba and Vanderweele's SACE estimand formula (1) are generated.

  ○ For the ith subject with $U_i = ss$ (always PP), we generate independent normal potential outcomes Y_1 and Y_0 as

$$Y_{1i} \sim N\left(50 + \delta, 30^2\right),$$
$$Y_{0i} \sim N\left(50, 30^2\right).$$

Independent Bernoulli potential outcomes Y_1 and Y_0 are generated as

$$Y_{1i} \sim Bernoulli(0.4 + \delta),$$
$$Y_{0i} \sim Bernoulli(0.4).$$

Here, δ is the true value of the SACE estimand. For normal potential outcomes, δ is the mean difference between TEST and RLD; for Bernoulli potential outcomes, δ is the proportion difference between TEST and RLD.

○ For the ith subject with $U_i = s\bar{s}$ (PP with TEST only), we generate independent normal and binary potential outcomes of Y_1 under TEST:
$Y_{1i} \sim N(50 + \delta + \beta_1, \ 30^2)$ and

$$Y_{1i} \sim Bernoulli(0.4 + \delta + \beta_1),$$

where $\beta_1 = E(Y_{1i}|U = s\bar{s}) - E(Y_{1i}|U = ss)$.

○ Likewise, for the ith subject with $U_i = \bar{s}s$ (PP with RLD only), we generate independent normal and binary potential outcomes of Y_0 under RLD as
$Y_{0i} \sim N(50 + \beta_0, \ 30^2)$ and

$$Y_{0i} \sim Bernoulli(0.4 + \beta_0),$$

where $\beta_0 = E(Y_{0i}|U = \bar{s}s) - E(Y_{0i}|U = ss)$

○ Generate the observed outcome Y: For the ith subject, if $A_i = 1$ and $S_i = s$, then the observed outcome, Y_i, is the potential outcome of this subject under TEST, i.e., $Y_i = Y_{1i}$; if $A_i = 0$ and $S_i = s$, then the observed outcome, Y_i, is the potential outcome of this subject under RLD, i.e., $Y_i = Y_{0i}$; the outcome of Y when $S = \bar{s}$ is not defined as it has been previously explained.

Finite sample statistics

Suppose the equivalence hypothesis for the normally distributed potential outcomes is

$$H_0: \ \mu_T - \mu_R \leq -10 \ or \ \mu_T - \mu_R \geq 10 \ vs. \ H_1: -10 < \mu_T - \mu_R < 10,$$

where ± 10 are the predetermined BE margins. We reject the null hypothesis if the 90% CI of $\mu_T - \mu_R$ is contained within $[-10, 10]$. Similarly, suppose the equivalence hypothesis for the binary potential outcome is

$$H_0: P_T - P_R \leq -0.2 \ or \ P_T - P_R \geq 0.2 \ vs. \ H_1: -0.2 < P_T - P_R < 0.2,$$

where ± 0.2 are the predetermined margins. We reject the null hypothesis if the 90% Wald CI with Yates' continuity correction is contained within $[-0.2, 0.2]$.

We compare three different estimators for the SACE estimand (4) in the finite samples for randomized, parallel trials.

(1) The SACE estimator: The difference of the average potential outcomes between those always PP who are randomly assigned to TEST and those always PP who are randomly assigned to RLD. In a completely randomized trial, this is an unbiased estimate for the SACE estimand (4):

$$\hat{\delta}_{SACE} = \frac{\sum_{i=1}^n I(A_i = 1, U_i = ss)Y_{1i}}{\sum_{i=1}^n I(A_i = 1, \ U_i = ss)} - \frac{\sum_{i=1}^n I(A_i = 0, \ U_i = ss)Y_{0i}}{\sum_{i=1}^n I(A_i = 0, \ U_i = ss)}. \tag{9}$$

(2) The PP estimator (the estimator for the observed PP estimand (2)): The difference of the average observed outcomes between those observed PP in TEST and observed PP in RLD:

$$\hat{\delta}_{PP} = \frac{\sum_{i=1}^n I(A_i = 1, \ S_i = s)Y_i}{\sum_{i=1}^n I(A_i = 1, \ S_i = s)} - \frac{\sum_{i=1}^n (A_i = 0, \ S_i = s)Y_i}{\sum_{i=1}^n I(A_i = 0, \ S_i = s)}. \tag{10}$$

(3) The sensitivity estimator for SACE: The PP estimator (9) plus the bias component from Chiba and Vanderweele's formula (7) with the true values of sensitivity parameters β_0, β_1, and $\pi_{\bar{S}S}$ (known in the simulation) and unbiased estimators for p_0 and p_1 from the observed data.

$$\hat{\delta}_{SACE^s} = \hat{\delta}_{PP} + \frac{\pi_{\bar{S}S}}{\hat{p}_0}\beta_0 - \frac{\hat{p}_1 - \hat{p}_0 + \pi_{\bar{S}S}}{\hat{p}_1}\beta_1 \tag{11}$$

(4) The SACE estimator (9) is based on the original definition for SACE (4). However, in simulation, we know who is always PP but in real studies, this information is unknown. The PP estimator (10) is what is used in current practice. The PP estimator is unbiased for the observed PP estimand (2), but it can be biased for the SACE estimand (4). By using the sensitivity estimator (11), we can examine whether the sensitivity analysis approach can provide a reasonable estimator for SACE (4) in finite samples when we know the true values of the sensitivity parameters in simulation.

Simulation scenarios

We let the true effect size, δ, be -10 (at the BE margin of H_0), 0 (under a strong H_1), and -3 (under a weak H_1), and explore four different scenarios for each δ by changing the true values of β_0, β_1, and $\pi_{\bar{S}S}$.

- Scenario 1: No selection effects. We let $\beta_0 = \beta_1 = 0$ for both normal and binary potential outcomes, and let $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.7, 0.1, 0.1, 0.1)$.
- Scenario 2: Direct effects only ($\pi_{s\bar{s}} = \pi_{\bar{s}s} = 0$). We let $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.75, 0, 0, 0.25)$, and let $\beta_0 = \beta_1 = -4$ for normal potential outcomes and $\beta_0 = \beta_1 = -0.08$ for binary potential outcomes.
- Scenario 3: Fix principal strata proportions but vary selection effects. We fix the principal strata proportions at $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.7, 0.1, 0.1, 0.1)$. Then, let the selection effects $(\beta_0, \beta_1) = (-20, -8), (-10, -4), (-5, -3), (-2, -2), (2, 2), (5, 3), (10, 4), (20, 8)$ for normal potential outcomes, and let $(\beta_0, \beta_1) = (-0.2, -0.02), (-0.2, -0.08), (-0.1, -0.06), (-0.04, -0.04), (0.04, 0.04), (0.1, 0.6), (0.2, 0.08), (0.2, 0.02)$ for binary potential outcomes.
- Scenario 4: Fix selection effects but vary principal strata proportions. We fix the selection effects β_0 and β_1 at the value of -4 for normal potential outcomes and fix the selection effects β_0 and β_1 at the value of -0.08 for binary potential outcomes. We then fix the proportions of always PP (π_{ss}) and never PP ($\pi_{\bar{s}\bar{s}}$) at 0.7 and 0.1, respectively, and vary the proportion of PP with RLD $\pi_{\bar{s}s} = 0.2, 0.15, 0.12, 0.11, 0.09, 0.08, 0.05, 0$.

We use total sample sizes of 200, 400, 600, and 800 subjects with 1:1 allocation for TEST/RLD (N_T/N_R) and 100,000 simulated datasets for all scenarios.

Simulation results

For all estimators, we present the true SACE δ_{SACE} (4), the total number of subjects randomized to TEST or RLD (N_T or N_R), the number of subjects actually used in each estimator in TEST and RLD (n_T or n_r), the average estimate for the three estimators ($\hat{\delta}_{SACE}$ (9), $\hat{\delta}_{PP}$(10), $\hat{\delta}_{SACE^s}$ (11)), the average bias (Bias = $\hat{\delta} - \delta_{SACE}$), and the rejection rate RR (%), computed as the percent of replications in which the null hypothesis is rejected, i.e., the power under H_1 and Type 1 error under H_0.

Under Scenario 1 with no selection effects ($\beta_0 = \beta_1 = 0$), the simulation results are summarized in Table 1 (normal potential outcomes) and Table 3 (binary potential outcomes). Both the SACE

Table 1. Simulation results for normal potential outcomes under scenario 1: No selection effects; $(\beta_0, \beta_1) = (0, 0)$; $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.7, 0.1, 0.1, 0.1)$; Calculated bias $= 0$.

δ_{SACE}	N_T/N_R	$\hat{\delta}_{SACE}$[1]				$\hat{\delta}_{PP}$[2]				$\hat{\delta}_{SACE^s}$[3]			
		n_T/n_R	$\hat{\delta}_{SACE}$	Bias	RR (%)	n_T/n_R	$\hat{\delta}_{PP}$	Bias	RR (%)	n_T/n_R	$\hat{\delta}_{SACE^s}$	Bias	RR (%)
−10	100/100	70/70	−10.03	−0.03	3.81	80/80	−10.02	−0.02	4.45	80/80	−10.02	−0.02	4.45
	200/200	140/140	−9.98	0.02	5.06	160/160	−9.98	0.02	5.02	160/160	−9.98	0.02	5.02
	300/300	210/210	−10.02	−0.02	4.95	240/240	−10.02	−0.02	4.89	240/240	−10.02	−0.02	4.89
	400/400	280/280	−10.01	−0.01	4.91	320/320	−10.00	0.00	4.92	320/320	−10.00	0.00	4.92
0	100/100	70/70	0.00	0.00	24.27	80/80	0.00	0.00	34.46	80/80	0.00	0.00	34.46
	200/200	140/140	0.00	0.00	73.88	160/160	0.00	0.00	81.25	160/160	0.00	0.00	81.25
	300/300	210/210	0.00	0.00	92.06	240/240	0.00	0.00	95.41	240/240	0.00	0.00	95.41
	400/400	280/280	0.01	0.01	97.80	320/320	0.01	0.01	98.96	320/320	0.01	0.01	98.96
−3	100/100	70/70	−3.02	−0.02	20.44	80/80	−3.00	0.00	28.70	80/80	−3.00	0.00	28.70
	200/200	140/140	−2.98	0.02	59.17	160/160	−2.98	0.02	65.42	160/160	−2.98	0.02	65.42
	300/300	210/210	−3.01	−0.01	76.64	240/240	−3.00	0.00	81.69	240/240	−3.00	0.00	81.69
	400/400	280/280	−3.01	−0.01	86.51	320/320	−3.02	−0.02	90.28	320/320	−3.02	−0.02	90.28

[1] $\hat{\delta}_{SACE}$: the SACE estimator;
[2] $\hat{\delta}_{PP}$: the PP estimator;
[3] $\hat{\delta}_{SACE^s}$: the sensitivity estimator for SACE.

estimator, $\hat{\delta}_{SACE}$, and the PP estimator, $\hat{\delta}_{PP}$, are unbiased for the SACE estimand, δ_{SACE}. The sensitivity estimator for SACE, $\hat{\delta}_{SACE^s}$, is identical to that of the PP estimator, $\hat{\delta}_{PP}$, because the bias component (7) is zero under Scenario 1. Type 1 error (i.e., RR (%) under H_0 when $\delta = -10$ for normal potential outcomes, and $\delta = -0.2$ for binary potential outcomes) is generally controlled at the 5% level or lower for all estimators. The power of the PP estimator, $\hat{\delta}_{PP}$, is slightly higher than that of the SACE estimator, $\hat{\delta}_{SACE}$, because PP estimator $\hat{\delta}_{PP}$ includes not only always PP ($P(U = ss) = 70\%$), as included in SACE estimator $\hat{\delta}_{SACE}$ but also includes PP with TEST only that are randomly assigned to the TEST group ($P(U = s\bar{s}, A = 1) = 5\%$) and PP with RLD only randomly assigned the RLD group ($P(U = \bar{s}s, A = 0) = 5\%$), which results in a total of 80% of the subjects, as previously explained.

Under Scenario 2 with direct effects only (Table 2 for normal potential outcome and Table 4 for binary potential outcome), the proportions of PP with TEST only and PP with RLD only are zero ($\pi_{s\bar{s}} = \pi_{\bar{s}s} = 0$). Hence, the observed PP sample contains only the always PP. Therefore, all three estimators are identical, and they are unbiased for the SACE estimand, δ_{SACE}.

Table 2. Simulation results for normal potential outcomes under scenario 2: direct effects only; $(\beta_0, \beta_1) = (-4, -4)$; $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.7, 0, 0, 0.25)$; Calculated Bias $= 0$.

δ_{SACE}	N_T/N_R	$\hat{\delta}_{SACE}$[1]				$\hat{\delta}_{PP}$[2]				$\hat{\delta}_{SACE^s}$[3]			
		n_T/n_R	$\hat{\delta}_{SACE}$	Bias	RR (%)	n_T/n_R	$\hat{\delta}_{PP}$	Bias	RR (%)	n_T/n_R	$\hat{\delta}_{SACE^s}$	Bias	RR (%)
−10	100/100	75/75	−10.00	0.00	4.08	75/75	−10.00	0.00	4.08	75/75	−10.00	0.00	4.08
	200/200	150/150	−9.99	0.01	4.98	150/150	−9.99	0.01	4.98	150/150	−9.99	0.01	4.98
	300/300	225/225	−10.01	−0.01	4.98	225/225	−10.01	−0.01	4.98	225/225	−10.01	−0.01	4.98
	400/400	300/300	−10.00	0.00	5.01	300/300	−10.00	0.00	5.01	300/300	−10.00	0.00	5.01
0	100/100	75/75	0.05	0.05	29.36	75/75	0.05	0.05	29.36	75/75	0.05	0.05	29.36
	200/200	150/150	−0.01	−0.01	78.08	150/150	−0.01	−0.01	78.08	150/150	−0.01	−0.01	78.08
	300/300	225/225	0.00	0.00	93.99	225/225	0.00	0.00	93.99	225/225	0.00	0.00	93.99
	400/400	300/300	0.01	0.01	98.45	300/300	0.01	0.01	98.45	300/300	0.01	0.01	98.45
−3	100/100	75/75	−3.01	−0.01	24.85	75/75	−3.01	−0.01	24.85	75/75	−3.01	−0.01	24.85
	200/200	150/150	−3.01	−0.01	62.39	150/150	−3.01	−0.01	62.39	150/150	−3.01	−0.01	62.39
	300/300	225/225	−3.01	−0.01	79.20	225/225	−3.01	−0.01	79.20	225/225	−3.01	−0.01	79.20
	400/400	300/300	−2.98	0.02	88.68	300/300	−2.98	0.02	88.68	300/300	−2.98	0.02	88.68

[1] $\hat{\delta}_{SACE}$: the SACE estimator;
[2] $\hat{\delta}_{PP}$: the PP estimator;
[3] $\hat{\delta}_{SACE^s}$: the sensitivity estimator for SACE.

Table 3. Simulation results for binary potential outcomes under scenario 1: no selection effects; $(\beta_0, \beta_1) = (0, 0)$; $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.7, 0.1, 0.1, 0.1)$; Calculated bias $= 0$.

		$\hat{\delta}_{SACE}$[1]				$\hat{\delta}_{PP}$[2]				$\hat{\delta}_{SACE^s}$[3]			
δ_{SACE}	N_T/N_R	n_T/n_R	$\hat{\delta}_{SACE}$	Bias	RR (%)	n_T/n_R	$\hat{\delta}_{PP}$	Bias	RR (%)	n_T/n_R	$\hat{\delta}_{SACE^s}$	Bias	RR (%)
−0.2	100/100	70/70	−0.20	0.00	3.16	80/80	−0.20	0.00	3.25	80/80	−0.20	0.00	3.25
	200/200	140/140	−0.20	0.00	3.69	160/160	−0.20	0.00	3.73	160/160	−0.20	0.00	3.73
	300/300	210/210	−0.20	0.00	3.93	240/240	−0.20	0.00	4.04	240/240	−0.20	0.00	4.04
	400/400	280/280	−0.20	0.00	4.16	320/320	−0.20	0.00	4.27	320/320	−0.20	0.00	4.27
0	100/100	70/70	0.00	0.00	44.58	80/80	0.00	0.00	55.82	80/80	0.00	0.00	55.82
	200/200	140/140	0.00	0.00	89.93	160/160	0.00	0.00	94.12	160/160	0.00	0.00	94.12
	300/300	210/210	0.00	0.00	98.55	240/240	0.00	0.00	99.39	240/240	0.00	0.00	99.39
	400/400	280/280	0.00	0.00	99.80	320/320	0.00	0.00	99.93	320/320	0.00	0.00	99.93
−0.06	100/100	70/70	−0.06	0.00	37.26	80/80	−0.06	0.00	45.56	80/80	−0.06	0.00	45.56
	200/200	140/140	−0.06	0.00	74.38	160/160	−0.06	0.00	79.80	160/160	−0.06	0.00	79.80
	300/300	210/210	−0.06	0.00	89.24	240/240	−0.06	0.00	92.59	240/240	−0.06	0.00	92.59
	400/400	280/280	−0.06	0.00	95.63	320/320	−0.06	0.00	97.45	320/320	−0.06	0.00	97.45

[1] $\hat{\delta}_{SACE}$: the SACE estimator.
[2] $\hat{\delta}_{PP}$: the PP estimator.
[3] $\hat{\delta}_{SACE^s}$: the sensitivity estimator for SACE.

Table 4. Simulation results for binary potential outcomes under scenario 2: direct effects only; $(\beta_0, \beta_1) = (-0.08, -0.08)$; $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.7, 0, 0, 0.25)$; Calculated bias $= 0$.

		$\hat{\delta}_{SACE}$[1]				$\hat{\delta}_{PP}$[2]				$\hat{\delta}_{SACE^s}$[3]			
δ_{SACE}	N_T/N_R	n_T/n_R	$\hat{\delta}_{SACE}$	Bias	RR (%)	n_T/n_R	$\hat{\delta}_{PP}$	Bias	RR (%)	n_T/n_R	$\hat{\delta}_{SACE^s}$	Bias	RR (%)
−0.2	100/100	75/75	−0.20	0.00	3.36	75/75	−0.20	0.00	3.36	75/75	−0.20	0.00	3.36
	200/200	150/150	−0.20	0.00	3.74	150/150	−0.20	0.00	3.74	150/150	−0.20	0.00	3.74
	300/300	225/225	−0.20	0.00	3.95	225/225	−0.20	0.00	3.95	225/225	−0.20	0.00	3.95
	400/400	300/300	−0.20	0.00	4.08	300/300	−0.20	0.00	4.08	300/300	−0.20	0.00	4.08
0	100/100	75/75	0.00	0.00	50.72	75/75	0.00	0.00	50.72	75/75	0.00	0.00	50.72
	200/200	150/150	0.00	0.00	92.46	150/150	0.00	0.00	92.46	150/150	0.00	0.00	92.46
	300/300	225/225	0.00	0.00	99.09	225/225	0.00	0.00	99.09	225/225	0.00	0.00	99.09
	400/400	300/300	0.00	0.00	99.89	300/300	0.00	0.00	99.89	300/300	0.00	0.00	99.89
−0.06	100/100	75/75	−0.06	0.00	41.63	75/75	−0.06	0.00	41.63	75/75	−0.06	0.00	41.63
	200/200	150/150	−0.06	0.00	77.23	150/150	−0.06	0.00	77.23	150/150	−0.06	0.00	77.23
	300/300	225/225	−0.06	0.00	91.05	225/225	−0.06	0.00	91.05	225/225	−0.06	0.00	91.05
	400/400	300/300	−0.06	0.00	96.71	300/300	−0.06	0.00	96.71	300/300	−0.06	0.00	96.71

[1] $\hat{\delta}_{SACE}$: the SACE estimator;
[2] $\hat{\delta}_{PP}$: the PP estimator;
[3] $\hat{\delta}_{SACE^s}$: the sensitivity estimator for SACE.

In Scenario 3, we fix the principal strata proportions at $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.7, 0.1, 0.1, 0.1)$ and vary the paired values of the selection effects (β_0, β_1) from negative to positive. The upper panels (bias) of Figure 4 (for normal potential outcomes) and Figure 5 (for binary potential outcomes) show that the SACE estimator, $\hat{\delta}_{SACE}$, and the sensitivity estimator for SACE, $\hat{\delta}_{SACE^s}$ (knowing the true value of sensitivity parameters in the simulation), are both unbiased, whereas the PP estimator, $\hat{\delta}_{PP}$, is biased unless there are equal selection effects, i.e., $\beta_0 = \beta_1$ (and equal off-diagonal strata proportions $\pi_{\bar{s}s} = \pi_{s\bar{s}}$ as in this scenario). This is in line with previously introduced Condition 3: ideal-exact equality $(\beta_0 = \beta_1$ and $\pi_{s\bar{s}} = \pi_{\bar{s}s})$ for $\delta_{SACE} = \delta_{PP}$. However, the absolute value of bias increases as the selection effects between TEST and RLD become more unbalanced (i.e., $|\beta_1 - \beta_0|$ deviates from 0). All figures are identical in the upper panels of Figures 4 and 5 because bias is not a function of sample size or δ_{SACE}.

In the lower panels (rejection rate) of Figures 4 and 5, the Type 1 error rate (the first rows) is inflated with the PP estimator, $\hat{\delta}_{PP}$, as the selection effect becomes more unbalanced, and the inflation increases with the increase of sample size. On the contrary, the SACE estimator and the
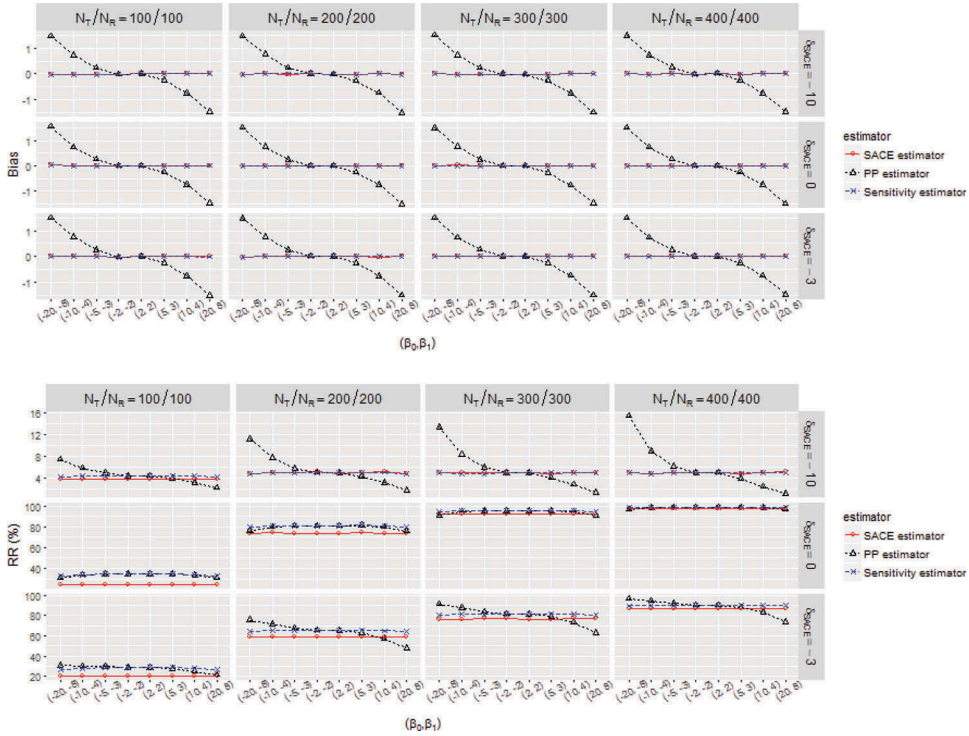
Figure 4. Bias (upper) and rejection rate (lower) for normal potential outcomes under scenario 3: Fix principal stratum proportions and vary selection effects. $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.7, 0.1, 0.1, 0.1)$. $(\beta_0, \beta_1) = (-20, -8), (-10, -4), (-5, -3), (-2, -2), (2, 2), (5, 3), (10, 4), (20, 8)$.

sensitivity estimator for SACE (with the true value of sensitivity parameters) both control the Type 1 error rate under 5%. When TEST and RLD have exactly the same average potential outcomes ($\delta_{SACE} = 0$, the second rows of the lower panels of Figures 4 and 5), the PP estimator, $\hat{\delta}_{PP}$, and the sensitivity estimator, $\hat{\delta}_{SACE^s}$ have about the same power, both slightly higher than that for the SACE estimator, $\hat{\delta}_{SACE}$, due to the inclusion of additional subjects as discussed previously (PP with TEST only assigned to TEST, and PP with RLD only assigned to RLD). However, under a weaker alternative ($\delta = -3$ or $\delta = -0.6$ for normal or binary outcomes, the last rows of the lower panels of Figures 4 and 5), the deviation in power for the PP estimator, $\hat{\delta}_{PP}$, from that of the SACE estimator, $\hat{\delta}_{SACE}$, and the sensitivity estimator for SACE, $\hat{\delta}_{SACE^s}$, becomes larger: inflated or deflated depending upon different combinations of selection effect.

In Scenario 4, we fix selection effects to be equal at −4 and −0.08 for normal and binary potential outcomes, respectively, and fix π_{ss} and $\pi_{\bar{s}\bar{s}}$ at 0.7 and 0.1, respectively. We vary $\pi_{\bar{s}s}$ from 0.2 to 0, so $\pi_{s\bar{s}}$ changes correspondingly from 0 to 0.2 (see Figure 6 for normal potential outcomes and Figure 7 for binary potential outcomes). The results are similar to those observed under Scenario 3. The only difference is that bias, inflation of Type 1 error, and change in power increase as the off-diagonal proportions become more unbalanced between PP with TEST only and PP with RLD only, rather than as the selection effect becomes more unbalanced between TEST and RLD as in Scenario 3. This happens when TEST and RLD have different safety and/or efficacy profiles. Our simulation results validate the three conditions we previously identified for $\delta_{SACE} = \delta_{PP}$. Simulation also demonstrates that the current PP estimator, $\hat{\delta}_{PP}$, is biased for the proposed causal SACE estimand, δ_{SACE}, under all other conditions except these three, and may inflate Type 1 error, and/or modify the power. In
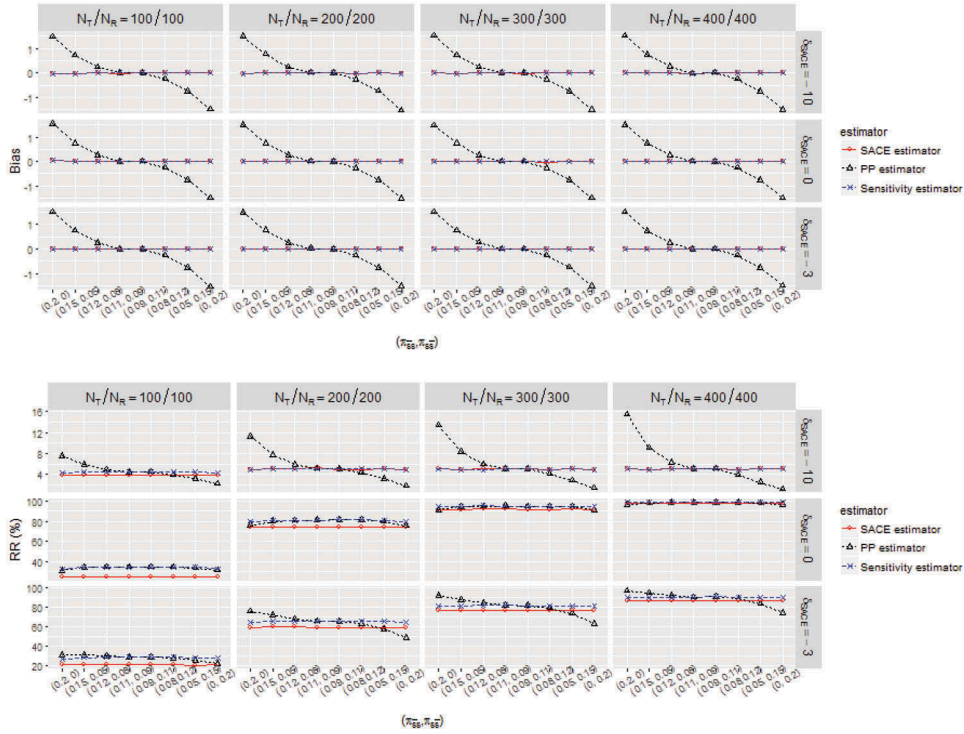
Figure 5. Bias (upper) and rejection rate (lower) for binary potential outcomes under scenario 3: Fix principal stratum proportions and vary selection effects. $(\pi_{ss}, \pi_{s\bar{s}}, \pi_{\bar{s}s}, \pi_{\bar{s}\bar{s}}) = (0.7, 0.1, 0.1, 0.1).(\beta_0, \beta_1) = (-0.2, -0.02), (-0.2, -0.08), (-0.1, -0.06), (-0.04, -0.04), (0.04, 0.04), (0.1, 0.6), (0.2, 0.08), (0.2, 0.02).$

particular, absolute bias increases as the off-diagonal proportions $(\pi_{s\bar{s}}, \pi_{\bar{s}s})$ or the selection effect (β_0, β_1) becomes more unbalanced, i.e., deviates from 0 from either direction.

Case study

We illustrate an estimation of the proposed co-primary causal estimands and how to apply the proposed tipping point sensitivity analysis by using clinical endpoint BE study data for acne vulgaris registered with ClinicalTrial.gov (ClinicalTrial.gov Identifier NCT01138514, available at *https://clinicaltrials.gov/ct2/show/NCT01138514*) (Table 5). This was a randomized, double-blind, multiple-site, placebo-controlled, parallel-group study on a total of 1555 (522:516:517 in TEST:RLD:Vehicle) men and women aged from 12 to 65 years with acne vulgaris. The primary endpoints were percent changes from baseline to 10 weeks in both inflammatory and non-inflammatory lesions, and the secondary endpoint was clinical success (yes or no) on Investigator's Global Assessment (IGA) at 10 weeks. In this paper, we will use the equivalence assessment of the secondary endpoint between TEST and RLD as an example because the equivalence hypothesis of the secondary endpoint is based on the difference of means, i.e., the difference of proportions in clinical success between treatment groups.

To establish equivalence of proportions in clinical success, we first need to evaluate the proposed co-primary causal estimand – proportion estimand δ_p – which can be unbiasedly estimated by the difference in the proportion of observed non-PP in a completely randomized trial – $\hat{\delta}_p$. Suppose the margin for the proposed proportion estimand, δ_p, is predetermined to be ± 15%, so that the hypothesis to be tested is $H_0 : \delta_p \leq -0.15 \ or \ \delta_p \geq 0.15$ vs. $H_1 : -0.15 < \delta_p < 0.15$. Table 5
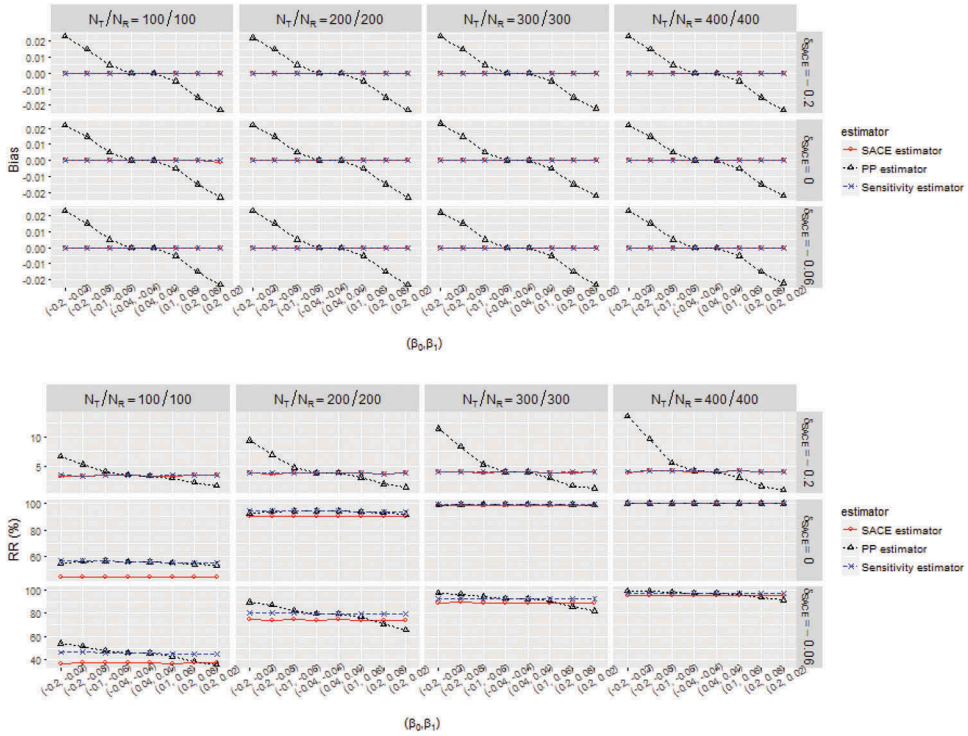
Figure 6. Bias (upper) and rejection rate (lower) for normal potential outcomes under scenario 4: Fix selection effects and vary principal stratum proportions; $(\beta_0, \beta_1) = (-4, -4)$; $\pi_{SS} = 0.7$; $\pi_{SS} = 0.1$; $\pi_{\bar{s}s} = 0.2, 0.15, 0.12, 0.11, 0.09, 0.08, 0.05, 0$.

shows that the proportion of observed non-PP subjects in TEST is $\hat{p}_1 = \Pr(S = \bar{s}|A = 1) = 17.6\%$ (92 out of 522 randomized subjects). In RLD, $\hat{p}_0 = \Pr(S = \bar{s}|A = 0) = 18.8\%$ (97 out of 516 randomized subjects). An unbiased estimator, $\hat{\delta}_p$, which estimates the difference in the proportion of observed PP between the two treatment arms, is -1.2%. The 90% CI for δ_p with Yates' continuity correction is $[-5.30\%, 2.96\%]$, which is contained within the prespecified equivalence margin of $[-15\%, 15\%]$. Based on this result, we conclude that the probability of being non-PP (or PP) is equivalent between TEST and RLD.

Next, we evaluate the equivalence in the SACE estimand, δ_{SACE}. As previously discussed, in the primary analysis, we will keep using the PP estimator, $\hat{\delta}_{PP}$. Among the observed PP population, the proportion of subjects with clinical success on IGA was 57.4% (247 out of 430) in TEST and 55.8% (234 out of 419) in RLD. The crude difference in success rate among the observed PP subjects between TEST and RLD (i.e., the PP estimator, $\hat{\delta}_{PP}$) is 1.6%, and the 90% Wald CI with Yates' continuity correction is $[-4.2\%, 4.7\%]$, which is contained within the prespecified equivalence limits of $[-20\%, 20\%]$.

We next evaluate the robustness of the primary analysis result based on the PP estimator, $\hat{\delta}_{PP}$, by the proposed tipping point sensitivity analysis for the SACE estimand, δ_{SACE}. As previously discussed, 82% of the observed PP subjects in the TEST group include not only the always PP subjects, but also the PP with TEST only subjects. However, we cannot distinguish who is always PP and who is PP with TEST only. For example, if 12% of subjects would be PP with RLD only, i.e., $\pi_{\bar{s}s} = 12\%$, then $\pi_{s\bar{s}} = \hat{p}_1 - \hat{p}_0 + \pi_{\bar{s}s} = 11\%$ of subjects would be PP with TEST only.

The estimated boundary of the proportion of RLD only $\pi_{\bar{s}s}$ is $[max(0, \hat{p}_0 - \hat{p}_1), min(\hat{p}_0, 1 - \hat{p}_1)] = [0, 0.176]$, meaning that the proportion of PP with RLD only ranges from 0% to 17.6%. Likewise, the proportion of PP with TEST only ranges from 1% to 18.6%. In
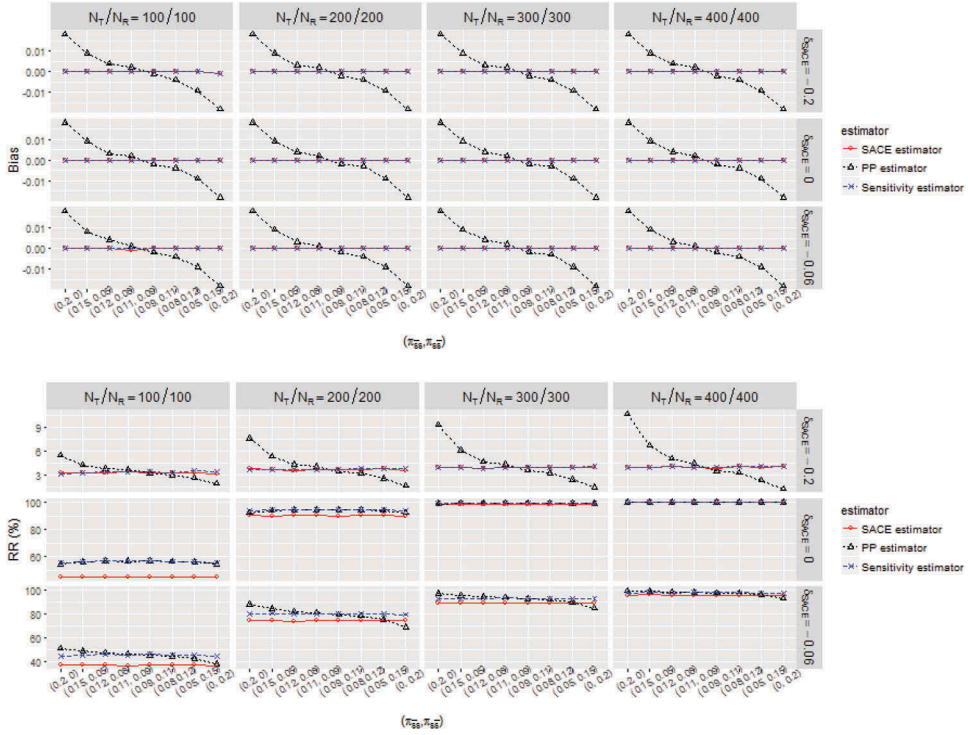
Figure 7. Bias (upper) and rejection rate (lower) for binary potential outcomes under scenario 4: Fix selection effects and vary principal stratum proportions; $(\beta_0, \beta_1) = (-0.08, -0.08)$; $\pi_{SS} = 0.7$; $\pi_{\bar{S}\bar{S}} = 0.1$; $\pi_{\bar{s}\bar{s}} = 0.2, 0.15, 0.12, 0.11, 0.09, 0.08, 0.05, 0$.

Figure 8, we chose a range of paired values of $(\pi_{\bar{s}s}, \pi_{s\bar{s}})$ within the boundary, i.e., $(0.04, 0.05), (0.08, 0.09), (0.12, 0.13), (0.16, 0.17)$, to illustrate the impact of different sensitivity parameters on the equivalence conclusion. We also vary the value of the selection effect (β_0, β_1) over a wide range of $[-50\%, 50\%]$ with an increment of 5%. For example, if $\beta_0 = 10\%$, this means that the proportion of potential clinical success with IGA is 10% higher among the stratum of PP with RLD only than in the stratum of always PP. On the contrary, if $\beta_1 = -10\%$, this means that the proportion of potential clinical success with IGA is 10% lower among the stratum of TEST only than in the stratum of always PP.

In each block of Figure 8, we fix the off-diagonal proportions $(\pi_{\bar{s}s}, \pi_{s\bar{s}})$ at one of the four selected paired values and plot the equivalence conclusion based on the sensitivity estimator for SACE $\hat{\delta}_{SACE^S}$ (green, reject H_0 or equivalence is established; red, fail to reject H_0 or fail to establish equivalence) corresponding to each paired value of selection effects (β_0, β_1) on the two axes. The equivalence conclusion based on the primary analysis was marked with a black "X," i.e., $\hat{\delta}_{PP} = 1.6\%$ in this example. Suppose $\pi_{\bar{s}s} = 12\%, \pi_{s\bar{s}} = 11\%, \beta_0 = 20\%$, and $\beta_1 = -5\%$, the sensitivity estimator for SACE $\hat{\delta}_{SACE^S}$ will shift the observed PP estimator from the primary analysis, $\hat{\delta}_{PP}$, by a bias component: $\hat{\delta}_{SACE^S} = \hat{\delta}_{PP} + \frac{\pi_{\bar{s}s}}{\hat{p}_0}\beta_0 - \frac{\hat{p}_1 - \hat{p}_0 + \pi_{\bar{s}s}}{\hat{p}_1}\beta_1 = 1.6 + 3.75 = 5.35$ with a 90% Wald CI of $[-0.44, 11.16]$. In Figure 8, we can see that the equivalence conclusion in this study is robust – passing equivalence in a majority of the sensitivity scenarios. The equivalence conclusion reverses only if the selection effects under TEST and RLD are very unbalanced (i.e., the absolute value of $\beta_0 - \beta_1$ is large). This happens if TEST and RLD have very different safety and efficacy profiles.

Table 5. Co-primary causal estimands for equivalence assessment using a clinical endpoint BE study.

	Hypothesis	TEST	RLD	TEST-RLD	90% CI ~
Randomized Subjects		522	516		
Observed Per-protocol (PP) subjects (S = s)		430	419		
Proposed Co-primary causal estimand 1: the proportion estimand (δ_p) comparing The proportion of potential non-PP between group $\delta_p = Pr(S_1 = \bar{s}) - Pr(S_0 = \bar{s})$, which is equal to the difference in the proportions of the observed non-PP: $Pr(S = \bar{s}|A = 1) - Pr(S = \bar{s}|A = 0)$	$H_0: \delta_p \leq -0.15$ $or\ \delta_p \geq 0.15;$ $H_1: -0.15 < \delta_p < 0.15$	$Pr(S = \bar{s}|A = 1)17.6\%$	$Pr(S = \bar{s}|A = 0)18.8\%$	$\hat{\delta}_p$-1.2%	90% CI [−5.3%, 2.96%] $\subset$ [−15%, 15%]
Endpoint (Y): Clinical success (yes or no) on the Investigator's Global Assessment (IGA)		(Yes) 183 (No)	234 (Yes) 185 (No)		
Proposed Co-primary causal estimand 2: the SACE estimand (δ_{SACE}) comparing average potential endpoints among the stratum of always PP between groups:$\delta_{SACE} = E(Y_1|U = ss) - E(Y_0|U = ss)$The primary analysis is based on PP estimator ($\hat{\delta}_{PP}$) comparing the average observed endpoints among observed PP between groups:$\delta_{PP} = E(Y|\ A = 1,\ S = s) - E(Y|\ A = 0,\ S = s)$	$H_0:\delta_{SACE} \leq -0.2\ or\ \delta_{SACE} \geq 0.2;$ $H_1: -0.2 < \delta_p < 0.2$	$E(Y|\ A = 1,\ S = s)57.4\%$	$E(Y|\ A = 0,\ S = s)55.8\%$	$\hat{\delta}_{PP}$1.6%	90% CI [−4.2%, 4.7%] $\subset$ [−20%, 20%]

*Data is registered at ClinicalTrial.gov Identifier NCT01138514 (available at https://clinicaltrials.gov/ct2/show/NCT01138514) for further information about this trial.
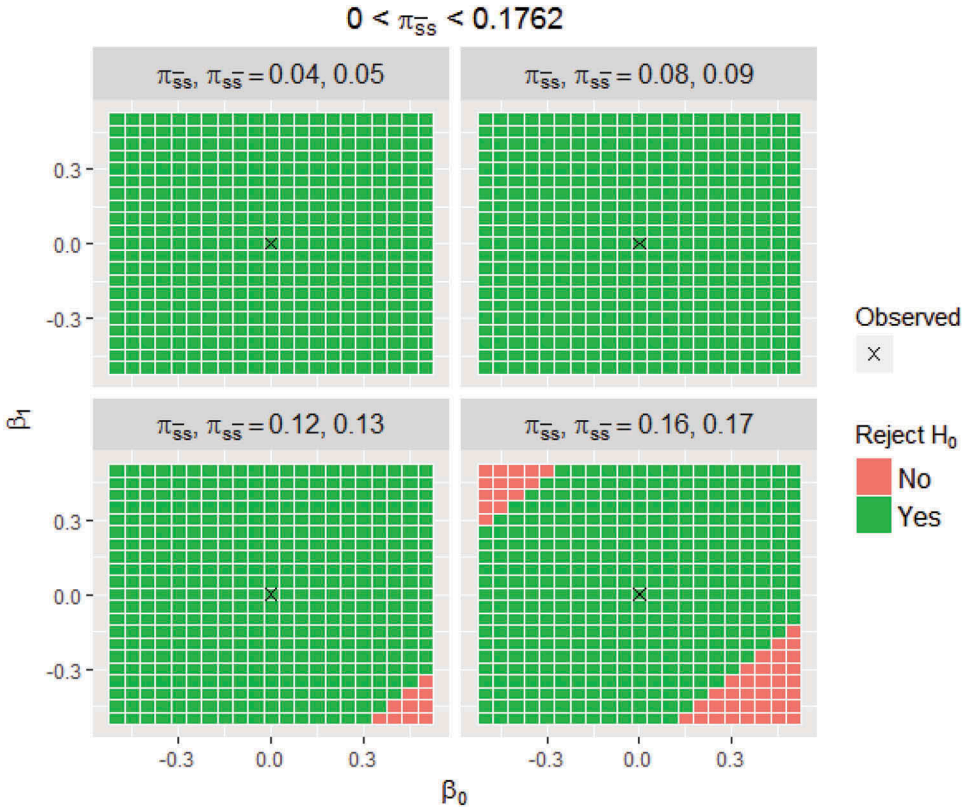~ Confidence interval (CI) is Wald CI with Yates' continuity correction.

Figure 8. Sensitivity analysis for the SACE estimand for number of participants with clinical success on IGA in a clinical endpoint BE study.

In conclusion, equivalence is established for the co-primary causal estimand, SACE, which compares the proportion of clinical success in the stratum of always PP. In addition, equivalence is also established in the proportion estimand, which compares the proportion of non-PP (or PP). In this paper, we use the difference in means as the primary measure. Next, we shall focus on the ratio of means.

Conclusions and discussion

In clinical endpoint BE studies, the current primary analysis for equivalence assessment is based on the observed PP population between TEST and RLD. However, observed PP status is a post-randomization intercurrent event; therefore, conditioning on it may break down randomization and introduce selection bias. The corresponding estimand of interest may not be causal (Frangakis and Rubin, 2002; International Conference on Harmonisation, 2017). Following the recommendation of the FDA Missing Data Working Group on using the "causal estimand of primary interest" (NRC, 2010; Permutt, 2016a), we introduce the framework of Frangakis and Rubin's principal stratification in causal inference to test equivalence in clinical endpoint BE studies with missing and noncompliance data, also recommended by the recently published ICH E9 (R1) addendum to address intercurrent events. To our knowledge, this is the first time causal inference has been applied in the assessment of equivalence.

We first propose four principal strata in the equivalence setting based on the joint potential PP status had a subject been assigned to TEST or RLD: "always PP," "PP with TEST only," "PP with

RLD only," and "never PP." This changes the original post-randomization intercurrent variable (i.e., the observed PP status) between treatment assignment and observed outcome to a pre-randomization variable (i.e., the potential PP status, similar to age and gender), so we can evaluate causal treatment effect on the potential outcome within the same pre-randomization principal stratum (i.e., a common set of units). We then propose co-primary causal estimands for the assessment of equivalence between TEST and RLD. First, the SACE estimand evaluates the "net" treatment effects in the stratum of "always PP," i.e., those who would comply and complete the study under both TEST and RLD. Second, the proportion estimand compares the proportion of being potential (or observed) non-PP (or PP) between the two treatment groups, which is a simple and indirect coverage of the other three strata. As previously discussed, oftentimes the status of PP or non-PP is related to the treatment effect, and therefore is an outcome itself. SACE is in line with the current observed PP estimand for equivalence testing by assessing the desired "net effect," but not based on the post-randomization "observed PP," which may not be a common set of units between the two treatment groups, but rather on the pre-randomization "potential PP" within a common set of units – the always PP stratum. Therefore, SACE is an improvement over the current observed PP estimand in that it is a principal stratification causal effect of primary interest as recommended by ICH E9 (R1) (ICH, 2017) and the FDA Missing Data Working Group (LaVange and Permutt, 2016; Permutt, 2016a). In addition, SACE excludes unobserved/missing outcomes or nonexistent outcomes, which simplifies the defining of the estimand by not imposing imputation for equivalence assessment.

We thoroughly evaluated the conditions under which the current PP estimand is equal to the causal SACE estimand and identified three conditions for which the current PP estimator is an unbiased estimate for SACE: 1) no selection effect ($\beta_0 = \beta_1 = 0$) : average potential outcome under TEST (or RLD) is the same regardless of whether a subject is PP with TEST (or RLD) only or is always PP; 2) direct effect only ($\pi_{s\bar{s}} = \pi_{\bar{s}s} = 0$) : no subjects are PP with TEST only or RLD only, i.e., being PP or non-PP is completely unrelated to the treatment; and 3) ideal-exact equality ($\beta_0 = \beta_1$ and $\pi_{s\bar{s}} = \pi_{\bar{s}s}$) : TEST and RLD have exactly the same safety and efficacy.

Our simulation results validate these three conditions and demonstrate that when none of these three conditions are satisfied, the current PP estimator is biased for the causal SACE estimand and may inflate the Type 1 error and/or change the power. In particular, bias increases as the off-diagonal proportions become more unbalanced (i.e., the absolute value of $\pi_{\bar{s}s}$ - $\pi_{s\bar{s}}$ is large) or the selection effects become more unbalanced between TEST and RLD (i.e., the absolute value of $\beta_0 - \beta_1$ is large). This happens if TEST and RLD have very different safety and efficacy profiles.

We propose to continue using the current PP estimator to estimate the SACE estimand, which assumes that at least one of the three identified conditions is satisfied. However, in conjunction with the PP estimator, we propose the use of a tipping point sensitivity analysis method to test the robustness of the primary analysis results by demonstrating the bias of the observed PP estimator for estimating the causal SACE estimand and the change to the equivalence conclusions when the underlying sensitivity parameters deviate from the three identified conditions but remain within clinically meaningful ranges. This sensitivity analysis approach has the following advantages. 1) It uses minimal causal assumptions, which are easily satisfied in a randomized clinical trial, rather than assuming untestable assumptions (e.g., ignorability or missing at random) that are hard to justify in clinical trials, or non-applicable assumptions for equivalence trials (e.g., monotonicity). 2) The formulas for the sensitivity estimators and the corresponding CIs are simple, and hence easy to implement in practice. 3) By plotting the equivalence decisions corresponding to specific values of sensitivity parameters, investigators can easily visualize how the conclusion may change when the three conditions are not satisfied, as well as the robustness of the primary analysis result. In particular, the clinical feasibility of the tipping point (the equivalence conclusion changes from passing to failing) needs to be evaluated by clinicians.

In summary, we introduce a principal stratification causal inference for the assessment of equivalence in clinical endpoint BE studies in the presence of intercurrent events, including missing

data and noncompliance, which was also recommended by the ICH E9 (R1) addendum and the FDA Missing Data Working Group. This will help the regulatory agencies better understand the potential impact of missing data and noncompliance data on equivalence decisions when the PP population is used to conduct the primary equivalence assessment. Future work will extend the results in this paper from the difference in means to the ratio in means as the measure of treatment effect. This work can also be applied to comparative clinical bio-similar and noninferiority studies.

Acknowledgments

Disclaimer

The views expressed in this article represent the opinions of the authors and do not represent the views and/or policies of the U.S. Food and Drug Administration.

References

Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434):444–455. doi:10.1080/01621459.1996.10476902

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics* 13(3):261–281. doi:10.2307/2527916

Chiba, Y. (2012). The large sample bounds on the principal strata effect with application to a prostate cancer prevention trial. *The International Journal of Biostatistics* 8(1):1–19. doi:10.1515/1557-4679.1365

Chiba, Y., VanderWeele, T. J. (2011). A simple method for principal strata effects when the outcome has been truncated due to death. *American Journal of Epidemiology* 173(7):745–751. doi:10.1093/aje/kwq418

National Research Council. (2010). The prevention and treatment of missing data in clinical trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Science and Education. Washington, DC, USA: The National Academies Press.

Cox, D. R. (1958). *Planning of Experiments*. Oxford, England: John Wiley & Sons.

Ding, P., Geng, Z., Yan, W., Zhou, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association* 106(496):1578–1591. doi:10.1198/jasa.2011.tm10265

Ding, P., Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B* 79(3):757–777. doi:10.1111/rssb.12191

Egleston, B. L., Scharfstein, D. O., Freeman, E. E., West, S. K. (2007). Causal inference for non-mortality outcomes in the presence of death. *Biostatistics (Oxford, England)* 8(3):526–545. doi:10.1093/biostatistics/kxl027

Food and Drug Administration 2010. Guidance for Industry: Non-inferiority clinical trials to establish effectiveness

Food and Drug Administration 2001. Guidance for Industry: Statistical approaches to establishing Bioequivalence

Frangakis, C. E., Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* 58(1):21–29. doi:10.1111/j.0006-341X.2002.00021.x

Frumento, P., Mealli, F., Pacini, B., Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association* 107 (498):450–466. doi:10.1080/01621459.2012.682532

Grosser, S., Park, M., Raney, S. G., Rantou, E. (2015). Determining equivalence for generic locally acting drug products. *Statistics in Biopharmaceutical Research* 7(4):337–345. doi:10.1080/19466315.2015.1093541

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research* 2(3):109–112. doi: 10.4103/2229-3485.83221

Hayden, D., Pauler, D. K., Schoenfeld, D. (2005). An estimator for treatment comparisons among survivors in randomized trials. *Biometrics* 61(1):305–310. doi:10.1111/biom.2005.61.issue-1

International Conference on Harmonisation (2017). ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. URL http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/08/WC500233916.pdf

Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with truncation-by-death. *Statistics & Probability Letters* 78(2):144–149. doi:10.1016/j.spl.2007.05.015

LaVange, L. M., Permutt, T. (2016). A regulatory perspective on missing data in the aftermath of the NRC report. *Statistics in Medicine* 35(17):2853–2864. doi:10.1002/sim.6840

Lawrence, X. Y., Li, B. V. (eds). (2014). *FDA Bioequivalence Standards*, Vol. 13. New York, USA: Springer.

Little, R. J., Long, Q., Lin, X. (2009). A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics* 65(2):640–649. doi:10.1111/j.1541-0420.2008.01066.x

Matilde Sanchez, M., Chen, X. (2006). Choosing the analysis population in non-inferiority studies: Per protocol or intent-to-treat. *Statistics in Medicine* 25(7):1169–1181. doi:10.1002/sim.2244

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated in 1990. *Statistical Science* 5(4):465–472. doi:10.1214/ss/1177012031

Permutt, T. (2016a). A taxonomy of estimands for regulatory clinical trials with discontinuations. *Statistics in Medicine* 35(17):2865–2875. doi:10.1002/sim.v35.17

Permutt, T. (2016b). Sensitivity analysis for missing data in regulatory submissions. *Statistics in Medicine* 35(17):2876–2879. doi:10.1002/sim.v35.17

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical Modelling* 7(9–12):1393–1512. doi:10.1016/0270-0255(86)90088-6

Robins, J. M., Hernán, M. A., Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)* 11(5):550–560.

Rosenbaum, P. R. (1984). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* 147(5):656–666. doi:10.2307/2981697

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688. doi:10.1037/h0037350

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* 6(1):34–58. doi:10.1214/aos/1176344064

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association* 75(371):591–593.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with" censoring" due to death. *Statistical Science* 21(3):299–309. doi:10.1214/088342306000000114

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Pharmacodynamics* 15(6):657–680. doi:10.1007/BF01068419

Snapinn, S. M. (2000). Noninferiority trials. *Current Controlled Trials in Cardiovascular Medicine* 1(1):19–21. doi:10.1186/cvm-1-1-019

Sun, W., Grosser, S., Tsong, Y. (2017). Ratio of means vs. difference of means as measures of superiority, noninferiority, and average bioequivalence. *Journal of Biopharmaceutical Statistics* 27(2):338–355. doi:10.1080/10543406.2016.1265536

Sun, W., Zhou, L., Grosser, S., Kim, C. (2016). A Meta-Analysis of missing data and non-compliance data in clinical endpoint bioequivalence studies. *Statistics in Biopharmaceutical Research* 8(3):334–344. doi:10.1080/19466315.2016.1201000

VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York, USA: Oxford University Press.

White, I. R. (2005). Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research* 14(4):327–347. doi:10.1191/0962280205sm406oa

Yang, F., Small, D. S. (2016). Using post-outcome measurement information in censoring-by-death problems. *Journal of the Royal Statistical Society: Series B* 78(1):299–318. doi:10.1111/rssb.12113

Zhang, J. L., Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by death. *Journal of Educational and Behavioral Statistics* 28(4):353–368. doi:10.3102/10769986028004353

Zhang, J. L., Rubin, D. B., Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association* 104(485):166–176. doi:10.1198/jasa.2009.0012