CLARIFYING IDENTIFICATION AND ESTIMATION OF TREATMENT EFFECTS IN THE SEQUENTIAL PARALLEL COMPARISON DESIGN

A PREPRINT

Benjamin Stockton[1], Michele Santacatterina[1], Soutrik Mandal[1], Charles M. Cleland[1], Erinn M. Hade[1], Nicholas Illenberger[1], Sharon Meropol[1], Andrea B. Troxel[1], Eva Petkova[1], Chang Yu[1], and Thaddeus Tarpey[1]

[1]Division of Biostatistics, Department of Population Health
New York University Grossman School of Medicine
New York, New York

November 26, 2025

ABSTRACT

Sequential parallel comparison design (SPCD) clinical trials aim to adjust active treatment effect estimates for placebo response to minimize the impact of placebo responders on the estimates. This is potentially accomplished using a two stage design by measuring treatment effects among all participants during the first stage, then classifying some placebo arm participants as placebo non-responders who will be re-randomized in the second stage. In this paper, we use causal inference tools to clarify under what assumptions treatment effects can be identified in SPCD trials and what effects the conventional estimators target at each stage of the SPCD trial. We further illustrate the highly influential impact of placebo response misclassification on the second stage estimate. We conclude that the conventional SPCD estimators do not target meaningful treatment effects.

***Keywords*—** clinical trials, latent variables, placebo response, estimand

1 Introduction

Placebo response complicates statistical inference and clinical understanding regarding treatment efficacy in placebo-controlled clinical trials in psychiatry and other fields[5, 1, 13]. As a result, studies have less ability than expected to detect effective treatments due to estimating a smaller active treatment effect. Estimating placebo response is one way to correct for this bias in statistical inference of average active treatment effects. The sequential parallel comparison design (SPCD) was developed to address this gap by attempting to identify the placebo non-responders during the first stage and conducting the second stage only among those identified as non-responders[5].

Typically participants are allocated to active and placebo arms at stage 1 using a 2:1 ratio with 2 participants assigned to placebo for each assigned to active treatment. A placebo response classification rule based on symptom improvement is then used to classify the placebo non-responders. Those classified non-responders are then randomized at stage 2 to active and placebo with roughly half assigned to each. In total this results in roughly a third of the participants receiving the active treatment at stage 1, another third receiving placebo and being categorized as placebo responders, and the final third being categorized as placebo non-responders and being re-randomized at stage 2. While alternative strategies to setting randomization allocation ratios are possible[14], the decision to choose a particular allocation is typically arbitrary. Additionally, the allocation scheme does not affect our results on the estimands and estimators in Section 3.

In the SPCD, participants are first randomized to active treatment or placebo at stage 1. After collecting outcomes for stage 1, participants in the placebo arm are classified as placebo non-responders or placebo responders. Placebo

Abbreviations: SPCD, Sequential Parallel Comparisons Design; ATE, Average Treatment Effect; NR, Non-Responder; PR, Placebo Responder.

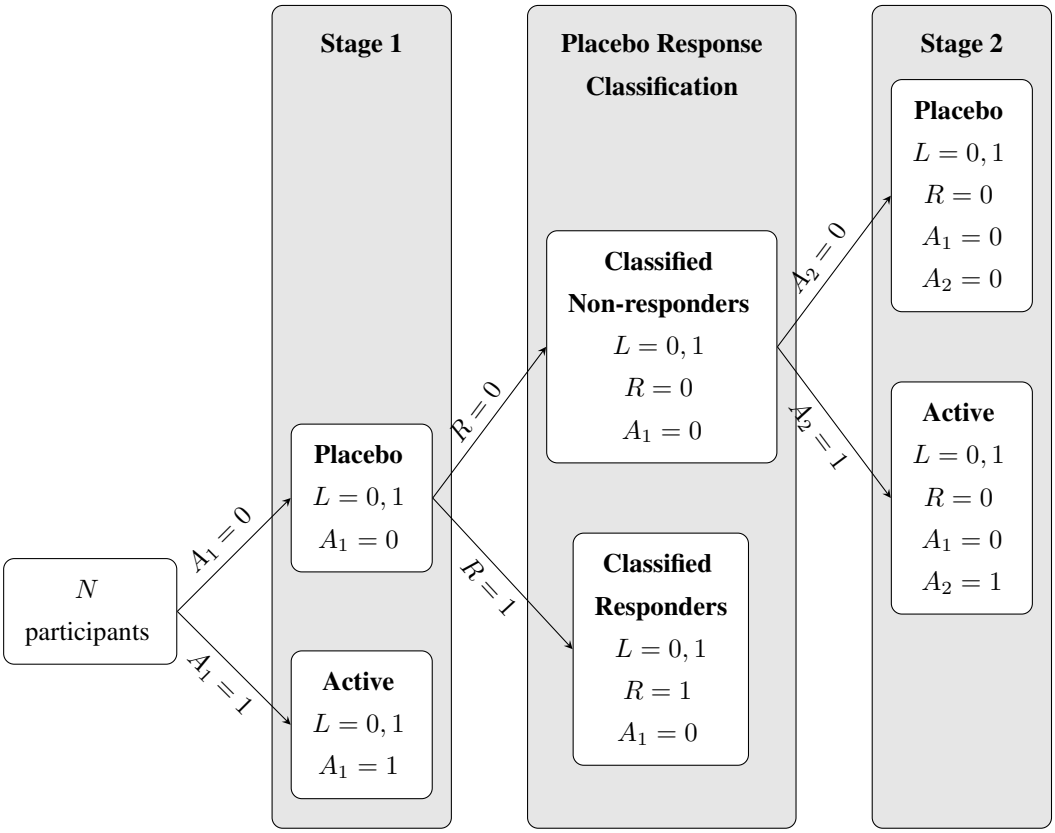

Figure 1: A diagram of the SPCD design. At stage 1, N participants are randomized to placebo ($A_1 = 0$) or active treatment ($A_1 = 1$). All participants are either latent true placebo responders ($L = 1$) or latent true placebo non-responders ($L = 0$), so each group is a mixture of placebo responders and non-responders. The placebo-treated participants are classified as placebo responders ($R = 1$) if their outcome is beyond a specified threshold and otherwise as placebo non-responders ($R = 0$). At stage 2, the placebo non-responders are randomized again and assigned to placebo ($A_2 = 0$) or active treatment ($A_2 = 1$). Classified placebo responders and those on active treatment from stage 1 continue on the originally assigned treatment without re-randomization.

non-responders are re-randomized in stage 2 to receive placebo or active treatment. Figure 1 displays the trial design. Note that in this paper we use the term estimand, treatment effect, causal effect, and causal parameter interchangeably.

The intuition behind the design is that the placebo response in stage 1 dilutes the measured active treatment effect while the effect measured at stage 2 should only capture the active treatment effect without placebo response. The literature on statistical analysis for the SPCD proposes an overall estimate resulting from a weighted average of the stage 1 and stage 2 active treatment effect estimates[1, 13, 14, 2, 3]. However, statistical inference under this design comes with significant complications[13, 14, 2, 12, 7]. In the rest of the manuscript, we use responders and non-responders to refer specifically to placebo response, unless otherwise noted. For brevity, we also use treatment effect or average treatment effect to refer specifically to the active treatment effect or average active treatment effect respectively.

Prior work[2] discussed six issues found with the conventional SPCD estimators. These issues include (1) placebo response biasing estimates, (2) assuming the active treatment effect is constant over time, (3) uncertainty in how to set the weight for the overall active treatment effect estimate averaging the stage 1 and stage 2 estimates, (4) unclear interpretation of the overall estimate, (5) inconsistency between the directions of the stage 1 and overall estimate, and (6) the common assumption that stage 1 and stage 2 are independent.

In this manuscript, we apply the counterfactual and causal inference framework[10] to illustrate the incoherence of the conventional weighted SPCD estimator, and to highlight an additional complication: placebo responder status misclassification. Others have discussed misclassification, primarily to motivate better predictive models[13, 12, 7]. Here we show how crucial accurate placebo response classification is to accurately estimating the active treatment effects in the SPCD.

$$
\begin{aligned}
Y_{i,0} &= g_{Y_0}(U_{i,Y_0}),\\
A_{i,1} &= g_{A_1}(U_{i,A_1}),\\
Y_{i,1} &= g_{Y_1}(Y_{i,0}, L_i, A_{i,1}, U_{i,Y_1}),\\
R_i &= C_R(Y_{i,0}, Y_{i,1}),\\
A_{i,2} &= \begin{cases} g_{A_2}(U_{i,A_2}) & \text{if } A_{i,1}=0,\ R_i=0,\\ 0 & \text{if } A_{i,1}=0,\ R_i=1,\\ 1 & \text{if } A_{i,1}=1, \end{cases}\\
Y_{i,2} &= g_{Y_2}(Y_{i,0}, L_i, A_{i,1}, A_{i,2}, U_{i,Y_2}).
\end{aligned}
$$

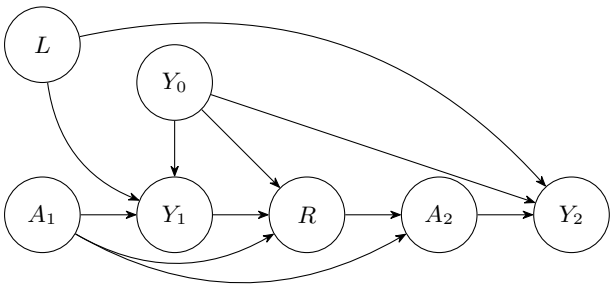

Figure 2: This causal DAG depicts the structural causal model underlying the SPCD with the addition of our latent placebo responder status L.

In Section 3, we develop quantities of interest, discuss identification assumptions required for estimation, and derive expectations for the conventional estimators in the SPCD using the causal inference framework. Using these derivations, we argue that the SPCD-specific estimands are not clearly defined and are either not targeted by the conventional estimators due to placebo response misclassification, or are targeted under unrealistic scenarios. Some of the unrealistic scenarios defeat the purpose of using an SPCD design (we provide more details on Sections 3.3.2, 3.3.3, and 5). The stage 1 estimand and estimator do target the average treatment effect for all participants, but this is achievable with a standard parallel randomized design. We demonstrate the importance of accurate placebo response classification using a simulation study in Section 4. Finally, in Section 5, we summarize our results, connect them to previously published work, and provide some direction on how to approach analysis for SPCD trials.

2 Background

In this section, we introduce some common notation to clarify the meanings of various estimands and estimators previously proposed in the literature.

2.1 SPCD Set Up

We consider the binary latent placebo responder framework proposed by Rybin et al.[13] wherein the population is a mixture of placebo responders and non-responders. A variation of the latent responder status is also used more generally as a counterfactual to categorize both drug response and placebo response[7]. Both latent responder approaches[13, 7] use an expectation maximization (EM) algorithm[4] to account for the unknown responder status when fitting their respective models.

2.1.1 Notation

For each of $i \in \{1, \ldots, n\}$ study participants, let L_i denote a binary latent variable indicating whether a participant would respond to placebo ($L_i = 1$) or not ($L_i = 0$). Let $Y_{i,0}$, $Y_{i,1}$, and $Y_{i,2}$ denote the participant's baseline (e.g., pre-treatment symptom severity), stage 1 and stage 2 outcomes, respectively. Let $A_{i,1}$ denote the randomized arm at Stage 1, i.e, placebo ($A_{i,1} = 0$) or active treatment ($A_{i,1} = 1$) and $A_{i,2}$ the assigned or randomized arm at stage 2, respectively. Finally, let R_i represent the placebo-responder indicator, i.e., participants are *classified* as responders ($R_i = 1$) or non-responders ($R_i = 0$). The observed data are $D = (Z_1, \ldots, Z_n)$, where Z_i represents the data for the i^{th} participant, i.e., $Z_i = (Y_{i,0}, Y_{i,1}, Y_{i,2}, R_i, A_{i,1}, A_{i,2})$ following some multivariate probability distribution ¶. The latent exogenous variables U_{Y_0}, U_{A_1}, U_{Y_1}, and U_{Y_2} are unmeasured and contain background information or influences on the observed variables[8].

2.1.2 A structural causal model and associated DAG

To clarify the SPCD set up, we use the structural causal model (SCM) and directed acyclic graph (DAG)[8] described in Figure 2. From the next paragraph, we omit the participant index i for better readability.

In this SCM, L, U_{Y_0}, U_{A_1}, U_{Y_1}, and U_{Y_2} are latent (exogenous) variables. The baseline symptom severity Y_0 depends only on exogenous variables U_L, and U_{Y_0}. In Stage 1, participants are randomized to placebo or active treatment following the randomization mechanism $g_{A_1}(\cdot)$. The Stage 1 outcome Y_1 depends on baseline symptom severity (Y_0), latent (true) placebo response status (L), treatment assignment (A_1), and exogenous factors. Placebo-treated

participants are *classified* as responders ($R = 1$) or non-responders ($R = 0$) using a classifier $R = C_R(Y_0, Y_1)$ based on baseline and stage 1 data (Y_0, Y_1). In Stage 2, placebo-treated participants *classified* as placebo non-responders are re-randomized following f_{A_2} to placebo or active treatment, while all others continue with their Stage 1 treatment, (A_2). The final outcome Y_2 depends on baseline severity (Y_0), latent response type (L), and both treatment assignments (A_1, A_2). Y_2 does not depend on R conditional on A_2 because R only determines whether stage 2 treatment (A_2) will be re-randomized (when $R = 0$ and $A_1 = 0$), otherwise $A_2 = A_1$. The stage 2 treatment (A_2) and latent placebo response status (L) then contain all the influence of R on Y_2.

It is important to note that, while L is unobserved, it is conceptually treated as a baseline variable in terms of causal temporality. It represents an inherent placebo response status that exists prior to treatment assignment and affects post-baseline outcomes, including Y_1 and Y_2. In contrast, the observed placebo response classification R is not a baseline variable—it is deterministically defined as a function of Y_0 and Y_1, and thus occurs *after* the outcome Y_1 is realized. From a causal ordering perspective, this places L upstream of all outcomes, while R is downstream of Y_1.

To put the latent and classified responders in context, consider pain score as an outcome. Conceptually, we consider a true placebo responder $(L = 1)$ to be a participant for whom the placebo treatment provides some therapeutic benefit compared to if they had not been treated at all. A true placebo non-responder $(L = 0)$ is someone whose potential pain score trajectory is unaffected by receiving the placebo. However, due to variance in the outcomes, it is plausible that a true placebo responder $(L = 1)$ may be misclassified as a non-responder $(R = 0)$ due to insufficient improvement to meet the responder threshold and vice versa for true placebo non-responders. As a result, we would very likely still have a mixture of true placebo responders and non-responders in the second stage of the SPCD, which would limit the ability to estimate the average active treatment effect without any placebo response.

3 SPCD Estimands and Estimators

We discuss potential estimands targeted in a SPCD in Section 3.1 as functions of counterfactual variables[8]. Then, we introduce assumptions for their identification in Section 3.2. Finally, we discuss conventional SPCD estimators in Section 3.3.

We assume all population average active treatment effects are time invariant, meaning the effect of the active treatment compared to placebo is the same at each stage for each group: the whole population (mixture of placebo responders and non-responders), the true placebo non-responders, and the true placebo responders. Let Δ_{all} be the average treatment effect in the population as a whole (Equation 4), Δ_{NR} be the average treatment effect for true placebo non-responders (Equation 2), Δ_{PR} be the average treatment effect for true placebo responders (Equation 3), and Δ_{placebo} be the average effect due to placebo response among placebo responders relative to the non-responders (Equation 1). Each of these will be defined and discussed in further detail in Section 3.1.

The stage 1 SPCD estimator is denoted $\hat{\theta}_1$ (Equation 9) and the stage 2 SPCD estimator for participants classified as placebo non-responders is $\hat{\theta}_2$ (Equation 11). The overall weighted estimator is a weighted average of the stage 1 and stage 2 estimators with pre-specified weight $0 \leq w \leq 1$; $\hat{\theta}_w = w\hat{\theta}_1 + (1 - w)\hat{\theta}_2$ (Equation 13). All three estimators are defined in detail in Section 3.3.

3.1 SPCD Estimands

Denote the counterfactual outcomes at times $t = 1, 2$ as $Y_t^{a,l}$ with the superscript treatment assignment a and latent placebo response status l. The latent placebo response is included in the counterfactual notation to emphasize that this is an unobserved property of the counterfactual outcome. Let $\mu_t^{a,l} = E(Y_t^{a,l})$ be the expected value at stage t of the outcome after treatment a with latent placebo status l, and $p_L = P(L = 1)$ be the prevalence of true placebo responders in the population.

Let us also assume that Y_0 follows a single baseline distribution without any effects due to treatment or placebo response by randomization. Then $E(Y_0|A_1, A_2) = E(Y_0) = \mu_0$.

Four estimands are summarized in Table 1 in terms of average effects (e.g., Average Treatment Effect, ATE): the average placebo response effect Δ_{placebo} given by Equation 1, the ATE for placebo non-responders Δ_{NR} by Equation 2, the ATE for placebo responders Δ_{PR} by Equation 3, and the ATE for all Δ_{all} by Equation 4.

$$\Delta_{\text{placebo}} = E(Y_1^{a_1=0,l=1}) - E(Y_1^{a_1=0,l=0}) \tag{1}$$

$$\Delta_{NR} = E(Y_1^{a_1=1}) - E(Y_1^{a_1=0,l=0}) \tag{2}$$

$$\Delta_{PR} = \Delta_{NR} - \Delta_{\text{placebo}} \tag{3}$$

$$\Delta_{\text{all}} = \Delta_{NR} - p_L \Delta_{\text{placebo}} \tag{4}$$

Most previous literature[1, 13, 14, 2, 3, 12] defines the average treatment effect for all participants Δ_{all} as the effect between treated participants and a mixture of placebo responders and non-responders. The treatment effects Δ_{NR} for the placebo non-responders, and Δ_{PR} for placebo responders are defined less often in the literature[13, 2]. We consider Δ_{PR} to be equivalent to $\Delta_{NR} - \Delta_{\text{placebo}}$. To our knowledge, previous work has not defined a quantity equivalent to Δ_{placebo}, the size of the placebo response for placebo responders relative to placebo non-responders.

Table 1 clarifies two important points: 1) estimands targeted by SPCD depend on unobserved non-responder and/or responder outcomes $Y_t^{a=0,l=0}$ and $Y_t^{a=0,l=1}$, for which clear and transparent assumptions are required [†], and 2) Δ_{all} and Δ_{NR} would be equivalent if the magnitude of any placebo response were zero ($\Delta_{\text{placebo}} = 0$) or if there were no placebo responders ($p_L = 0$). If either of these conditions were true, then the rationale for using the SPCD would vanish.

Table 1: Average effects as functions of counterfactual outcomes and latent placebo responder status. Average treatment effect is abbreviated as ATE.

Quantity	**Notation**	**Population Subset**
Average Placebo Response Effect	$\Delta_{\text{placebo}} = E(Y_1^{a_1=0,l=1}) - E(Y_1^{a_1=0,l=0})$	Stage 1 Placebo-treated Participants; $A_t = 0$
ATE for Placebo Non-Responders	$\Delta_{NR} = E(Y_1^{a_1=1}) - E(Y_1^{a_1=0,l=0})$	Placebo Non-responders; $L = 0$
ATE for Placebo Responders	$\Delta_{PR} = \Delta_{NR} - \Delta_{\text{placebo}}$	Placebo Responders; $L = 1$
ATE for All Participants	$\Delta_{\text{all}} = \Delta_{NR} - p_L \Delta_{\text{placebo}}$	All Participants

3.2 Identification Assumptions

We provide below causal and statistical assumptions needed to identify the estimands discussed in our previous section. For brevity, we define these assumptions for the Stage 1 counterfactual outcomes, but they can be naturally extended to Stage 2 as well.

Assumption 1 (Consistency and the Stable Unit Treatment Values Assumption (SUTVA)).

$$Y_1 = Y_1^{A_1,L} \quad \text{and} \quad Y_{i,1}^{a,l} \perp\!\!\!\perp A_{j,1} \text{ for all } i \neq j.$$

Assumption 1 states that each unit's observed outcome corresponds to their latent placebo response status and their counterfactual outcome under the treatment actually received and that treatment is consistently applied or given across individuals. In other words, there are not different versions of the placebo or active treatment due to variation in providers, dosage, adherence, etc. It also assumes no interference between units. Together these comprise the stable unit treatment values assumption[11, 9, 6]. This is encoded in the structural causal model (Figure 2).

Assumption 2 (Ignorability).

$$A_1 \perp\!\!\!\perp (Y_1^{1,1}, Y_1^{1,0}, Y_1^{0,1}, Y_1^{0,0})$$

Assumption 2 is satisfied by proper randomization and holds by design.

Assumption 3 (Positivity).

$$0 < P(A_1 = a) < 1 \quad \text{for all } \ a.$$

Assumption 3 ensures that all treatment arms are possible and holds by design.

At Stage 2, only placebo-treated participants classified as non-responders are re-assigned, so we need $P(A_2 = a | R_1 = 0, A_1 = 0) > 0$ for classified placebo non-responders and stage 2 treatment assignments $a = 0, 1$. If re-randomized, then this should also hold by design.

Assumption 4 (Latent-type stability).

$$L \perp\!\!\!\perp A_1$$

[†]Note how the placebo treated outcomes $Y_t^{a=0,l}$ can be expressed as a mixture between the unobserved non-responder outcome $Y_t^{a=0,l=0}$ and unobserved responder outcome $Y_t^{a=0,l=1}$, $Y_t^{a=0,l} = I(L=1)Y_t^{a=0,l=1} + (1 - I(L=1))Y_t^{a=0,l=0}$.

Assumption 4 asserts that the latent placebo response L is fixed at baseline and unaffected by treatment assignment. This can fail if the drug biologically alters placebo pathways (e.g., via neural priming), so that $L^{a_1=1} \neq L^{a_1=0}$. Then the mean outcome under placebo for placebo responders, $\mu_t^{a=0,l=1}$, no longer represents the counterfactual placebo outcome for the drug-treated group, breaking the validity of $\Delta_{\text{placebo}} = \mu_t^{a=0,l=1} - \mu_t^{a=0,l=0}$.

Assumption 5 (Effect homogeneity across L).

$$\mu_t^{a=1,l=1} = \mu_t^{a=1,l=0} \tag{5}$$

Assumption 5 posits that the active treatment effect is the same regardless of latent placebo response status.

If $\mu_t^{a=1,l=1} \neq \mu_t^{a=1,l=0}$, then we are unable to identify the ATE for non-responders ($L = 0$) because we do not classify active treatment participants as responders/non-responders in order to be able to estimate $\mu_t^{a=1,l=0}$. While this assumption could be violated in practice, we make it here for brevity in the exposition of the issues with the estimators. Alternative strategies that do not require this assumption have been proposed but are beyond the scope of this paper[7].

Using the latent placebo response variable, we can express the placebo outcome at stage 1 as the mixture $Y_1^{a=0,l} = I(L=1)Y_1^{a_1=0,l=1} + (1 - I(L=1))Y_1^{a_1=0,l=0}$, $I(\cdot)$ is an indicator function.

Assumption 6 (Mixture identifiability in the placebo arm (model-based)). Given that Y_1 follows a mixture, let $p_L = P(L = 1)$ be the unknown placebo responder prevalence, $f_{Y_1}(\cdot;\xi)$ be component densities of Y_1^l belonging to a parametric family $\mathcal{F} = \{f(\cdot;\xi)\}$, and ξ_l be parameters indexing the family $\mathcal{F}$ for latent response status $L = l$. Then the density of the stage 1 outcome in the placebo arm is,

$$f_{Y_1|A_1=0}(y) = p_L f_{Y_1}(y;\xi_1) + (1 - p_L) f_{Y_1}(y;\xi_0), \tag{6}$$

such that the mapping $(p_L, \xi_0, \xi_1) \mapsto f_{Y_1|A=0}(\cdot)$ is one-to-one within that model family, and the mixture is identified under that parametric model.

Assumption 6 is needed to identify $\mu_1^{0,l} = E(Y_1^{0,l})$ for $l = 0, 1$ from the observed Stage-1 placebo data. If $\xi_1 \approx \xi_0$ or $p_L \approx 0$, the mixture model is weakly identified.

$$E(Y_1 \mid A_1 = 0) \;=\; \int y\, f_{Y_1|A_1=0}(y)\, dy \;=\; p_L\, \mu_1^{0,1} + (1 - p_L)\, \mu_1^{0,0}, \tag{7}$$

To illustrate, consider the average placebo effect $\Delta_{\text{placebo}} = \mu_1^{0,1} - \mu_1^{0,0}$ and the latent placebo responder status L. Let $f_{L=0}(y)$ and $f_{L=1}(y)$ be nonparametric component densities of the stage 1 placebo outcome under $L = 0$ and $L = 1$, respectively. From the data we observe only the placebo-arm mixture:

$$f_{Y_1|A_1=0}(y) \;=\; p_L\, f_{L=1}(y) + (1 - p_L)\, f_{L=0}(y). \tag{8}$$

Then, nonparametrically, the triple $(p_L, f_{L=0}, f_{L=1})$ is *not* point identified from a single observed mixture $f_{Y_1|A_1=0}$, because many different triples can generate the same mixture. Hence $\mu_1^{0,0}$ and $\mu_1^{0,1}$ (and therefore Δ_{placebo}) are not nonparametrically identified.

Assumption 6 imposes a parametric mixture structure, as in Equation 6 which enables *model-based* identification: if the mapping $(p_L, \xi_0, \xi_1) \;\longmapsto\; f_{Y_1|A_1=0}(\cdot)$ is one-to-one within the chosen family $\mathcal{F}$, then (p_L, ξ_0, ξ_1) and thus $(\mu_1^{0,0}, \mu_1^{0,1})$ are identified *under that model*. Note this identifies component means, but does not imply we can correctly classify individual L's. While other identification strategies for causal inference with latent variables have been proposed (see, e.g., [16, 15]), in this manuscript we adopt a model-based approach. As our goal is to clarify causal effects in SPCD trials, the model-based framework suffices for that purpose, as shown in the following sections and in the Supplementary Materials.

We also make one additional *statistical* assumption of stationarity for the average effects.

Assumption 7 (Stationarity). $E(\Delta_{g,1}) = E(\Delta_{g,2}) = \Delta_g$ and $Var(\Delta_{g,1}) = Var(\Delta_{g,2})$ where $\Delta_{g,t}$ is the average effect for sub-group $g \in \{\text{all}, NR, PR, \text{placebo}\}$ at stage $t = 1, 2$.

Assumption 7 assumes that the effects remain constant over time and between stages. This form of stationarity is necessary for statistical inference based on the conventional SPCD estimators.

Note that we do not assume that $E(Y_1^{a_1=0,l=0}) = E(Y_0)$ as this assumption is unlikely to hold, e.g., in psychiatric or pain treatment trials where we may expect conditions to improve over time regardless of treatment or placebo effects. Both psychiatric and pain treatments are common applications of the SPCD.

Based on these assumptions, we provide identification of Table 1 estimands in Appendix 5 and that will be used to clarify what estimands are targeted by the conventional SPCD estimators.

3.3 The Conventional SPCD Estimators

Let the total N participants be partitioned into N_A participants assigned to active treatment in stage 1, and N_P participants assigned to placebo in stage 1. Let N_{NR} and N_{PR} be the random number of classified placebo non-responders and responders respectively, $N_{NR} + N_{PR} = N_P$; $N_{NR} = \sum_{i=1}^{N_P} I(R = 0)$ for some placebo response indicator R; and let n_{NR} and n_{PR} be the observed counts of N_{NR} and N_{PR} respectively. The number of non-responders N_{NR} is further split into N_{PP} and N_{PA} for placebo-placebo assignment and placebo-active assignment with observed counts n_{PP} and n_{PA} respectively, such that $N_{PP} + N_{PA} = N_{NR}$ and $n_{PP} + n_{PA} = n_{NR}$.

The commonly used weighted estimator[1, 13, 14, 2, 3, 7] combines stage 1, $\hat{\theta}_1$, and stage 2, $\hat{\theta}_2$, to get $\hat{\theta}_w$ using an pre-determined scalar weight w. See Appendix 5 for full derivations of the expectations.

3.3.1 Stage 1 Estimator (All Participants)

The stage 1 estimator $\hat{\theta}_1$ is defined as in Equation 9,

$$\hat{\theta}_1 = \frac{1}{n_A}\sum_{i=1}^{n}(Y_{i,1} - Y_{i,0})I(A_{i,1} = 1) - \frac{1}{n_P}\sum_{i=1}^{n}(Y_{i,1} - Y_{i,0})I(A_{i,1} = 0), \tag{9}$$

which ideally corresponds to the ATE of the population Δ_{all} at the end of stage 1.

Under Assumptions 1-7, its expectation is as given in Equation 10.

$$\begin{aligned} E(\hat{\theta}_1) &= \Delta_{NR} - p_L\Delta_{\text{placebo}} \\ &= \Delta_{\text{all}}. \end{aligned} \tag{10}$$

Note that we do not condition on R or L in this expectation calculation. The placebo response is marginalized out by the mixture $Y_1^{a=0,l} = I(L = 1)Y_1^{a_1=0,l=1} + (1 - I(L = 1))Y_1^{a_1=0,l=0}$. So we can expect this estimator to be unbiased for Δ_{all} in all scenarios.

3.3.2 Stage 2 Estimator (Non-Responders)

Given the number of placebo non-responders $n_{NR} = n_{PP} + n_{PA}$, the stage 2 estimator can be defined as in Equation 11,

$$\begin{aligned} \hat{\theta}_2 = &\frac{1}{n_{PA}}\sum_{i=1}^{n}(Y_{i,2} - Y_{i,1})I(A_{i,1} = 0, R_i = 0, A_{i,2} = 1) \\ &- \frac{1}{n_{PP}}\sum_{i=1}^{n}(Y_{i,2} - Y_{i,1})I(A_{i,1} = 0, R_i = 0, A_{i,2} = 0). \end{aligned} \tag{11}$$

We would like for $\hat{\theta}_2$ to unbiasedly estimate Δ_{NR}. Under Assumptions 1-7 (modified accordingly for Stage 2), $E(\hat{\theta}_2)$ is derived as in Equation 12,

$$\begin{aligned} E(\hat{\theta}_2) &= \Delta_{NR} - P(L = 1|R = 0)\Delta_{\text{placebo}} \\ &\neq \Delta_{NR}. \end{aligned} \tag{12}$$

$E(\hat{\theta}_2)$ is neither the ATE for all participants (Δ_{all}) nor the ATE for the placebo non-responders (Δ_{NR}). This is because $R \neq L$, so that the population in stage 2 is still some mix of true responders and non-responders, which in turn suppresses the estimate of Δ_{NR} as we will see in the simulation. If we want $E(\hat{\theta}_2) = \Delta_{NR}$, then we would need $P(L = 1|R = 0) = 0$ meaning the classifier $C_R(y_0, y_1)$ must have perfect negative predictive power. Alternatively, if $P(L = 1|R = 0) = P(L = 1)$, then $E(\hat{\theta}_2) = \Delta_{NR} - P(L = 1)\Delta_{\text{placebo}} = \Delta_{\text{all}}$, but that would imply that R is independent of L. (This may happen if $C_R(y_0, y_1) = 0$ for all y_0, y_1, i.e. everyone is classified as a placebo non-responder.) Note that we cannot evaluate the predictive capabilities of $C_R(y_0, y_1)$ in practice because L is latent.

3.3.3 Overall Weighted Estimator

Finally, the overall estimator $\hat{\theta}_w$ is a weighted average of $\hat{\theta}_1$ and $\hat{\theta}_2$, given by Equation 13,

$$\hat{\theta}_w = w\hat{\theta}_1 + (1 - w)\hat{\theta}_2. \tag{13}$$

Under Assumptions 1-7,

$$\begin{aligned} E(\hat{\theta}_w) &= wE(\hat{\theta}_1) + (1 - w)E(\hat{\theta}_2) \\ &= \Delta_{NR} - \Delta_{\text{placebo}} \left[wP(L = 1) + (1 - w)P(L = 1|R = 0) \right]. \end{aligned} \tag{14}$$

Like $E(\hat{\theta}_2)$, $E(\hat{\theta}_w)$ is also ill-defined in that it targets neither Δ_{all} nor Δ_{NR}, as well as mixing the marginal and conditional probabilities of latent placebo response. The estimator $\hat{\theta}_w$ will be unbiased if $\Delta_{\text{all}} = \Delta_{NR}$. But this is only true when there are only placebo non-responders so that the estimands coincide, e.g., when $\Delta_{\text{placebo}} = 0$ or $P(L = 1) = 0$. The quantities Δ_{placebo}, $P(L = 1|R = 0)$, and $P(L = 1)$ are unobservable and cannot be directly estimated because $L = l$ is latent, so these conditions cannot be checked.

These results hold generally for the conventional SPCD estimators, however we have not specified two other crucial analysis decisions; (i) how to choose w to actually weight the combination, and (ii) how to choose the classification function $C_R(y_0, y_1)$.

A similar latent responder status framework could be used for active treatment response such that participants may or may not respond to the active treatment. However, active treatment non-responders are then effectively receiving a placebo and so are part of the placebo responder status framework as well. This is a key part of the latent responder status framework[7].

3.4 The (Lack of) Attention Given to Classification

A limitation of previous work on SPCD is the absence of substantial discussion of placebo response classification, with limited exceptions[13, 12, 7]. Typically, classification is done by setting a cut-off or threshold c as in Equation 15,

$$C_R(y_0, y_1) = I(y_1 - y_0 > c) \text{ or } C_R(y_0, y_1) = I(y_1 > c), \tag{15}$$

so that participants with large improvements over baseline or large outcomes in the placebo arm at Stage 1 are labelled as placebo responders. As previously noted[13, 12, 7], this approach is subjective due to the arbitrary choice of c, error-prone, and often not supported by any relevant clinical data. However, given that the placebo response status is unobservable, this is a relatively intractable classification problem.

One attempt[12] at addressing the misclassification issue proposed using a re-randomization design that re-assigns treatment to all Stage 1 placebo participants rather than just placebo non-responders and then weights their response in Stage 2 by a measure of their responder status. Another approach[13] went further in the discussion of misclassification and proposed using a normal mixture model to classify the responders and non-responders.

This mixture model approach was expanded upon by Liu et al. (2022)[7] and the misclassification issue was given more extensive discussion. They note perfect classification as being a crucial assumption in the derivation of their strata-level effects, estimators, and inference. Liu et al.[7] also criticize the use of thresholds as subjective and error-prone which can potentially bias the inference. Instead they also propose fitting a mixture model with an EM algorithm and using the fitted probabilities of belonging to each strata given the data to then sample the "observed" responder statuses $S_i^{a_1}$ and $S_i^{a_2}$ for each individual i. However, the sampled responder statuses are still treated as known. They discuss how an approach that treats these statuses as stochastic may be an avenue for further development. The

stochastic approach is beyond the scope of this paper. Liu et al.'s simulation studies[7] demonstrate that the causal estimators derived through their approach are biased when assuming the sampled placebo response are fixed. This is likely due to the misclassification which they note in their discussion.

4 Simulation Study

In order to illustrate the bias of $\hat{\theta}_2$ and $\hat{\theta}_w$ for Δ_{NR} and Δ_{all} respectively, we devised a simulation study to evaluate the impacts of the size of the placebo response Δ_{placebo} (also called the average placebo effect or APE) and error standard deviation σ_ϵ on the misclassification of placebo non-response and the resulting bias in the estimators.

In these simulations, we assume that the baseline Y_0 is independent of treatment assignments A_1, A_2 and latent response L, that the average treatment effect for non-responders Δ_{NR} is equal at stage 1 and 2, that the average placebo effect Δ_{placebo} is also equal at stage 1 and 2, and that the residual standard deviation σ_ϵ is constant from baseline ($t = 0$) to stage 2 ($t = 2$) and equal in the placebo and active treatment groups conditional on the latent placebo response and previous outcome observation.

Our goal is to evaluate the inferential ability of each of the estimators $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_w$ with $w = 0.5$ to estimate the two ATEs of interest: Δ_{all} and Δ_{NR}. We are interested in the impact of the size of the average placebo effect (APE or Δ_{placebo}) and the variance in the outcome σ^2_Y based on the assumption that these two quantities have a strong influence on the classification ability of $C_R(y_0, y_1)$. If σ_ϵ is large, then there is high variance in the outcomes which will create significant overlap between placebo responders and non-responders making classification difficult. If Δ_{placebo} is large, then the gap between Δ_{all} and Δ_{NR} is larger and it should be easier to distinguish between the placebo responders and non-responders.

The data for this simulation experiment are generated as follows:

- $Y_0 \sim N(0, \sigma^2_\epsilon)$
- $L \sim Binom(1, 0.5)$
- Treatment is assigned at random with 1/3 of participants receiving active treatment $A_1 = 1$ and the remaining receiving placebo $A_1 = 0$.
- $Y_1 = Y_0 + \Delta_{NR}A_1 + \Delta_{\text{placebo}}L(1 - A_1) + \epsilon_1$ where $\epsilon_1 \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$.
- $R = I(Y_1 - Y_0 \geq q_{p_r})$ where q_{p_r} is the p_rth quantile of the differences $Y_1 - Y_0$ among the placebo group.
  - The classifier is $C_R(y_0, y_1) = I(y_1 - y_0 \geq q_{p_r})$.
- Phase 2 treatment is assigned randomly to the non-responders $R_1 = 0$ with half receiving active treatment ($A_2 = 1$) and half receiving placebo ($A_2 = 0$).
- $Y_2 = Y_1 + \Delta_{NR}A_2 + \Delta_{\text{placebo}}L(1 - A_2) + \epsilon_2$ where $\epsilon_2 \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$.

Note that the true placebo response status is determined by the latent factor L and not the observed responder indicator R and that we assume that responder type L is the same at all time points.

The estimators are calculated as described in Section 3.3. The Stage 2 estimator $\hat{\theta}_2$ and weighted estimator $\hat{\theta}_w$ are calculated using two classifiers; first, thresholding as described above is used to classify placebo response, and second, using an oracle revealing the true placebo responder status L. The thresholding approach can be interpreted as a conventional estimate that would be reported in practice. The oracle approach can be interpreted as an implausible best case scenario, where we have full information on the responder status. The simulations support our theoretical results in Section 3 that even under these ideal conditions, the weighted estimate that has been popularized throughout the SPCD literature will always be biased for either well-defined quantity Δ_{all} or Δ_{NR} as long as there is a placebo effect.

In the simulations, we hold weight w fixed at $w = 0.5$ for simplicity and to illustrate how the weighted estimator does not target either Δ_{NR} nor Δ_{all}. The response threshold is fixed at the empirical 50th percentile of the placebo arm's responses $q_{0.5}$. $\Delta_{\text{all}} = 0$ is also fixed as the null to represent the scenario where the treatment provides no benefits on average over the population mixture of placebo response and no placebo response. The placebo response is varied from the null $\Delta_{\text{placebo}} = 0$ to a relatively high placebo response $\Delta_{\text{placebo}} = 1$ with uncertainty in Y varied from $\sigma_\epsilon = 0.001$ to $\sigma_\epsilon = 10$. 10,000 Monte Carlo samples were generated for each simulation setting.

In Figure 3, we display the average negative predictive values (NPV; $P(L = 0|R = 0)$) from the simulation study for the threshold classifier described in Section 4. The lowest possible NPV without knowledge of L is 0.5 in our

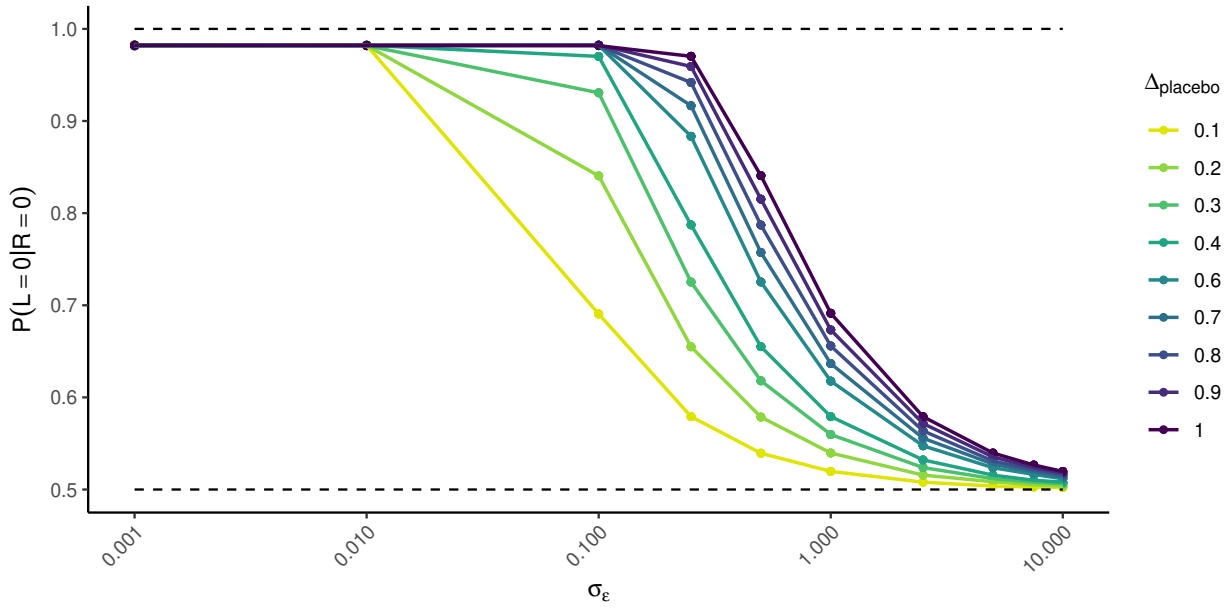

Figure 3: The average negative predictive values $P(L = 0|R = 0)$ for the threshold classifier used in the simulation study.

simulations since we set exactly half of participants to be placebo non-responders. A NPV of 0.5 then corresponds to assigning responder classifications at random (or equivalently calling everyone a placebo responder). A high NPV corresponds to correctly identifying the true placebo non-responders and only including those participants in the stage 2 re-randomization.

We can see that when Δ_{placebo} is small (the light lines and points) the NPV begins dropping towards the baseline 0.5 for lower values of σ_ϵ than for larger Δ_{placebo}. This indicates that non-responder classification is easier when there is either a medium to large average placebo effect or small residual standard deviation relative to the average placebo effect. Both of these conditions result in clear separation between the modes of the mixture of placebo responders and non-responders.

In Figure 4, the residual standard deviation σ_ϵ is displayed along the x-axis. The y-axis displays the bias in estimating Δ_{all} in Figure 4a and the bias in estimating Δ_{NR} in Figure 4b. The lines and points are color coded by the average placebo effect Δ_{placebo} from 0 to 1 varied by increments of 0.1 with lower values in lighter colors and higher values in darker colors. The sub-panels correspond to the various estimators. "Stage 1" ($\hat{\theta}_1$) measures the difference for the stage 1 treatment arms while "Stage 2" ($\hat{\theta}_2$) measures the difference for the stage 2 classified placebo non-responders and "Weighted" ($\hat{\theta}_w$) is the weighted average of these. The "Stage 2 (Oracle)" and "Weighted (Oracle)" are the stage 2 and weighted estimators applied to the true placebo non-responders. A horizontal black dashed line marks zero bias in each sub-panel. Deviations from the horizontal line at zero indicate bias in estimating the respective ATEs.

From Figure 4a we can see that $\hat{\theta}_1$ is consistently able to unbiasedly estimate Δ_{all}. For moderately large standard deviations ($\sigma_\epsilon > 1$) that exceed the average placebo effect size, the non-oracle estimators $\hat{\theta}_2$ and $\hat{\theta}_w$ tend towards Δ_{all} as σ_ϵ increases because the classifier cannot correctly distinguish between responders and non-responders even for large Δ_{placebo} as evidenced by Figure 3.

In Figure 4b, the stage 2 oracle estimator $\hat{\theta}_2^{\text{oracle}}$ is unbiased for Δ_{NR} as long as σ_ϵ is small (< 5), while the non-oracle $\hat{\theta}_2$ is only unbiased for Δ_{NR} when $\sigma_\epsilon < 1$. The weighted estimator, $\hat{\theta}_w$, fails to estimate either Δ_{NR} or Δ_{all} when $\Delta_{\text{placebo}} > 0$ and $\sigma_\epsilon > 1$, and $\hat{\theta}_w^{\text{oracle}}$ is always biased for both Δ_{NR} and Δ_{all}.

Crucially, the weighted estimator $\hat{\theta}_w$ never estimates Δ_{NR} and is unbiased for either ATE only in the scenario where the SPCD provides no benefit over a conventional parallel design because the classifier is independent of true responder status $P(L = 1|R = 0) = P(L = 1)$. The weighted oracle estimator $\hat{\theta}_w^{\text{oracle}}$ is always biased for both ATE as long

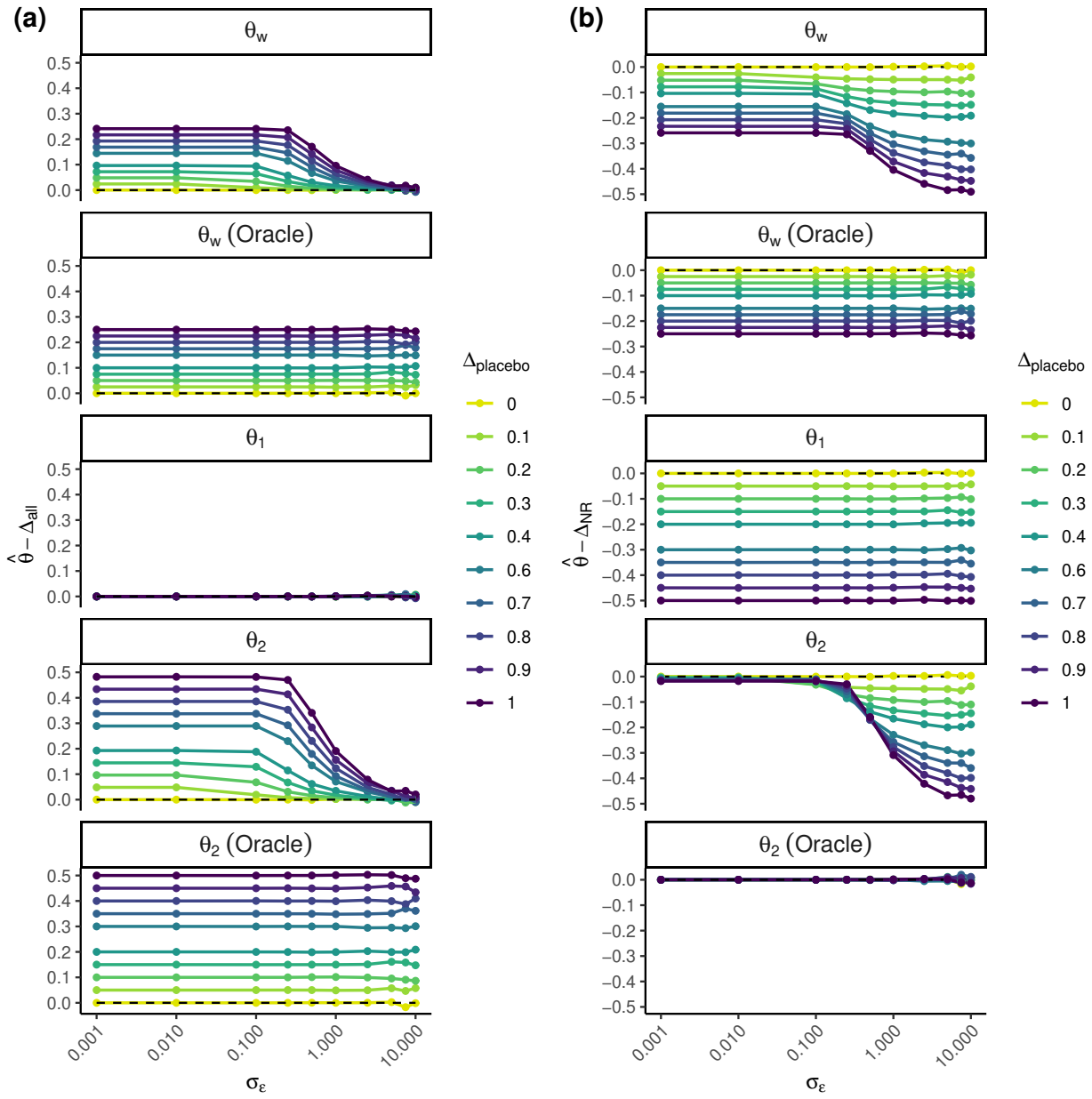

Figure 4: Simulation under the null; $\Delta_{\text{all}} = 0$. Varying the average placebo effect (APE) Δ_{placebo} and residual standard deviation σ_σ. The residual standard deviation is plotted along the x-axis. The y-axis in (a) is the estimated bias in estimating Δ_{all} for each estimator. The y-axis in (b) is the estimated bias in estimating Δ_{NR} for each estimator. The lines and points are color-coded by the pre-set Δ_{placebo}.

as $\Delta_{\text{placebo}} > 0$. When $\Delta_{\text{placebo}} = 0$, the estimands Δ_{NR} and Δ_{all} coincide and the estimators also coincide in expectation, so they are all unbiased in this scenario.

5 Discussion

In summary, we derived the estimands for each subset of participants (all participants, placebo non-responders, and placebo responders) in the SPCD while characterizing the placebo non-responder status as a latent binary variable. The latent status is estimated by classifier $C_R(y_0, y_1)$ using the baseline and stage 1 data for the placebo arm. We demonstrate that two of the conventional estimators for the SPCD, $\hat{\theta}_2$ (Equation 11) and $\hat{\theta}_w$, (Equation 13) do not target their intended estimands while $\hat{\theta}_1$ does unbiasedly estimate Δ_{all}. Our analytical and simulation results indicate that placebo non-response misclassification is a central cause of the bias.

From Equation 12, the conditions for which $E(\hat{\theta}_2) = \Delta_{NR}$, the desired expectation, are (i) $P(L = 1|R = 0) = 0$, equivalently, $R_i = 0 \Rightarrow L_i = 0$ for all placebo non-responders (i.e. the classifier only identifies true non-responders as non-responders), or (ii) $\Delta_{\text{placebo}} = 0$. However, both of these conditions are unverifiable. In the first condition (i) the probability of placebo responder misclassification, $P(L = 1|R = 0)$, depends on the unknown latent status L, so we cannot check the negative predictive performance of $C_R(y_1, y_0)$. The second condition (ii) requires that the placebo has no effect. This condition is not testable, is not typically a reasonable assumption, and defeats the purpose of using an SPCD to estimate the placebo effect. Condition (i), the assumption of perfect classification, is typically assumed implicitly in order to obtain an unbiased estimator without acknowledging that it is impossible to check if it is even plausible[13, 2, 3, 12].

The weighted estimator, $\hat{\theta}_w$, has been the primary inferential tool in the SPCD literature[1, 13, 14, 2, 12], despite the conceptual difficulty in interpreting its expectation given by Equation 14. Primarily, it is unclear how to interpret the weighted average component $(wP(L = 1) + (1 - w)P(L = 1|R = 0))$ of the expectation of $\hat{\theta}_w$. This conceptual challenge can be resolved under the above implausible conditions for unbiasedness for $\hat{\theta}_2$ or under the third condition (iii) that $P(L = 1|R = 0) = P(L = 1)$.

Under condition (iii), the expectation simplifies to $E(\hat{\theta}_w) = \Delta_{NR} - P(L = 1)\Delta_{\text{placebo}} = \Delta_{\text{all}}$. However, $P(L = 1|R = 0) = P(L = 1)$ occurs if $R \perp L$, $P(L = 1) = 0$, $\Delta_{\text{placebo}} = 0$, or $w = 1$ (see Result 9). All four conditions defeat the purpose of using the SPCD. The classified placebo responder status R being independent of L implies that we do not gain any information by performing the classification and stage 2 is essentially a replication of stage 1 because true placebo responders are not excluded from stage 2. The second and third conditions ($P(L = 1) = 0$ or Δ_{placebo}) imply that there are no placebo responders in the population to begin with. Lastly, choosing $w = 1$ results in only using the stage 1 data which reduces the SPCD to a simple randomized controlled trial.

If we are trying to estimate Δ_{NR}, then we should use $w = 0 \Rightarrow \hat{\theta}_w = \hat{\theta}_2$ and hope that $P(L = 1|R = 0)$ is small (see Result 10). When $w = 0$, the estimation only uses placebo arm stage 1 data to determine $C_R(y_0, y_1)$ and throws away the entire active treatment arm where $A_1 = 1$. In order to accurately estimate Δ_{NR} with $\hat{\theta}_2$, whichever (deterministic or stochastic) classifier $C_R(y_0, y_1)$ that is used to determine responder status R must perfectly identify true responders $L = 1$. If that is not the case then some true responders $L = 1$ may be classified as non-responders by $R = 0$ creating a mixture at Stage 2.

Because the true latent statuses are unknown, there is uncertainty in the classifications. In practice, we should require the inferential model or estimators to account for the classification uncertainty in the variance estimate of the stage 2 estimator. However, the classification uncertainty is unaccounted for in previous definitions of the SPCD estimator[2, 3, 7]. The previously shown suboptimal control of Type I error by the weighted SPCD estimator $\hat{\theta}_w$ may be partially explained by the stage 2 estimator failing to unbiasedly estimate Δ_{NR} due to the misclassification[3, 7].

The overall weighted estimator $\hat{\theta}_w$ is particularly problematic. Using any $w \neq 0, 1$ results in a biased weighted estimator $\hat{\theta}_w$ whose expectation (Equation 14) does not correspond to a quantity of clear clinical interest. The placebo non-responder ATE is adjusted for placebo response Δ_{placebo} by some weighted average of the proportion of placebo responders in the population and the proportion of responders that are misclassified as non-responders. It does not seem likely that this quantity would be of clinical interest or importance given that it is a function of the statistical analysis through the classifier $C_R(y_0, y_1)$.

Moreover, it is still unclear how to interpret what the combination of the SPCD stage 1 and stage 2 estimators measure when Δ_{all} and Δ_{NR} are not equal, even in the scenarios where we can unbiasedly estimate Δ_{all} or Δ_{NR}. We have not identified an explanation and description of this estimator, grounded in clinical meaning, of the weighted average of a marginal and conditional probability for the placebo non-response mean.

Our results are related to prior work[7], but highlight the importance of the accuracy of classification and provide a quantification of the bias of estimates due to misclassification. Where the previous work[7] uses a counterfactual variable to define the responder status so that both drug and placebo responders and non-responders are defined resulting in four strata, we define only placebo response through our non-counterfactual, latent status L which assumes that everyone would be a drug responder. This in effect collapses their never-responder and drug-only responder strata[7], while the always-responder strata correspond to our placebo responders and active treatment participants.

Additionally, Liu et al.[7] focus on the estimation of the individual strata effects at each time point, and the effects for never-responders and drug-only responders by taking a weighted average of their respective estimates at each stage. Their results assume that classification is accurate following the EM algorithm's maximum likelihood estimation and sampling of responder status. Our results provide insight of the impact of misclassification on the estimates.

Beyond the scope of this investigation, it may be better to model placebo response as a continuous latent variable rather than a binary indicator since it is plausible that everyone can exhibit some degree of placebo response along a continuum. While the binary responder status is a simplifying assumption that ties estimation to the well-known finite mixture model, it may not be a reasonable assumption for many of the applications in which SPCD are deployed, e.g., to study interventions for pain or psychiatric conditions.

Funding Statement

This research was supported in part by Grant Number 5U24NS113844-0 from the National Institute of Neurologic Disorders and Stroke and by NYU CTSA grant (UL1TR001445) from the National Center for Advancing Translational Sciences, National Institutes of Health.

This research is based upon work supported by the National Science Foundation under Grant No 2306556 and National Institute of Health Grant No 1R01AI197146-01.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

References

[1] Yeh-Fong Chen et al. "Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials". In: *Contemporary Clinical Trials* 32.4 (July 1, 2011), pp. 592–604. DOI: 10.1016/j.cct.2011.04.006. URL: https://www.sciencedirect.com/science/article/pii/S1551714411000966.

[2] George Y.H. Chi et al. "On clinical trials with a high placebo response rate". In: *Contemporary Clinical Trials Communications* 2 (Nov. 18, 2015). PMID: 29736445 PMCID: PMC5935859, pp. 34–53. DOI: 10.1016/j.conctc.2015.10.002. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5935859/.

[3] Yifan Cui, Semhar Ogbagaber, and H.M. James Hung. "Statistical inference problems in sequential parallel comparison design". In: *Journal of Biopharmaceutical Statistics* 29.6 (Nov. 2, 2019). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10543406.2019.1609014 PMID: 31035859, pp. 1116–1129. DOI: 10.1080/10543406.2019.1609014. URL: https://doi.org/10.1080/10543406.2019.1609014.

[4] A. P. Dempster, Martin Schatzoff, and Nanny Wermuth. "A Simulation Study of Alternatives to Ordinary Least Squares". In: *Journal of the American Statistical Association* 72.357 (1977). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 77–91. DOI: 10.2307/2286909. URL: http://www.jstor.org/stable/2286909.

[5] Maurizio Fava et al. "The Problem of the Placebo Response in Clinical Trials for Psychiatric Disorders: Culprits, Possible Remedies, and a Novel Study Design Approach". In: *Psychotherapy and Psychosomatics* 72.3 (2003), pp. 115–127. DOI: 10.1159/000069738. URL: https://karger.com/article/doi/10.1159/000069738.

[6] Miguel A Hernan and James M Robins. *Causal Inference: What If*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2020.

[7] Xiaoyan Liu et al. "Assessing treatment benefit in the presence of placebo response using the sequential parallel comparison design". In: *Statistics in Medicine* 41.12 (2022). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9349, pp. 2166–2190. DOI: 10.1002/sim.9349. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9349.

[8] Judea Pearl. *Causality*. Cambridge university press, 2009.

[9] James M. Robins, Miguel Ángel Hernán, and Babette Brumback. "Marginal Structural Models and Causal Inference in Epidemiology". In: *Epidemiology* 11.5 (Sept. 2000), pp. 550–560. DOI: 10.1097/00001648-200009000-00011. URL: http://journals.lww.com/00001648-200009000-00011.

[10] Donald B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66.5 (1974). Place: US Publisher: American Psychological Association, pp. 688–701. DOI: 10.1037/h0037350.

[11] Donald B. Rubin. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment". In: *Journal of the American Statistical Association* 75.371 (1980). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 591–593. DOI: 10.2307/2287653. URL: https://www.jstor.org/stable/2287653.

[12] Denis Rybin et al. "Placebo non-response measure in sequential parallel comparison design studies". In: *Statistics in Medicine* 34.15 (2015). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6494, pp. 2281–2293. DOI: 10.1002/sim.6494. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6494.

[13] Denis Rybin et al. "Placebo Response as a Latent Characteristic: Application to Analysis of Sequential Parallel Comparison Design Studies". In: *Journal of the American Statistical Association* 113.524 (Oct. 2, 2018). Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2017.1375930, pp. 1411–1430. DOI: 10.1080/01621459.2017.1375930. URL: https://doi.org/10.1080/01621459.2017.1375930.

[14] Roy N. Tamura, Xiaohong Huang, and Dennis D. Boos. "Estimation of treatment effect for the sequential parallel design". In: *Statistics in Medicine* 30.30 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4412, pp. 3496–3506. DOI: 10.1002/sim.4412. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4412.

[15] Junzhe Zhang and Elias Bareinboim. "Bounding Causal Effects on Continuous Outcome". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.13 (May 2021), pp. 12207–12215. DOI: 10.1609/aaai.v35i13.17449. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17449.

[16] Yaochen Zhu et al. "Causal Inference with Latent Variables: Recent Advances and Future Prospectives". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '24 (2024), pp. 6677–6687. DOI: 10.1145/3637528.3671450. URL: https://doi.org/10.1145/3637528.3671450.

Estimand Derivations

We provide identification results for each of the stage-1 estimands under the structural causal model described in Section 2, assuming Assumptions 1–6.

Result 1 (Average Placebo Effect). The average effect due to placebo is

$$\Delta_{\text{placebo}} = E(Y_1^{a_1=0,l=1}) - E(Y_1^{a_1=0,l=0}).$$

Proof. First, we establish the identification of the expectation of the counterfactuals from the observed data. For $l \in \{0, 1\}$,

$$\begin{aligned} E(Y_1^{a_1=0,l}) &= E(Y_1^{a_1=0,l} \mid A_1 = 0) && \text{by Assumption 2} \\ &= E(Y_1^{a_1=0,l} \mid A_1 = 0, L = l) && \text{by Assumption 4} \\ &= E(Y_1 \mid A_1 = 0, L = l) && \text{by Assumptions 1, 3} \\ &\text{identified from } f_{Y_1|A_1=0}(y) && \text{by Assumption 6} \end{aligned}$$

Then,

$$\begin{aligned} E(Y_1^{a_1=0,l=1} - Y_0) - E(Y_1^{a_1=0,l=0} - Y_0) &= \Big[E(Y_1^{a_1=0,l=1}) - E(Y_0)\Big] - \Big[E(Y_1^{a_1=0,l=0}) - E(Y_0)\Big] \\ &= E(Y_1^{a_1=0,l=1}) - E(Y_1^{a_1=0,l=0}) \\ &= \Delta_{\text{placebo}}. \end{aligned} \tag{16}$$

Note that $E(Y^{a_1=0,l=1}) = \Delta_{\text{placebo}} + E(Y^{a_1=0,l=0})$.

Result 2 (Average Treatment Effect for Placebo Non-Responders). The average treatment effect for true placebo non-responders is

$$\Delta_{\text{NR}} = E[Y_1^{a_1=1,l=0}] - E[Y_1^{a_1=0,l=0}].$$

Proof. Again, we begin by identifying the counterfactual's expectations from the data.

$$\begin{aligned} E(Y_1^{1,0}) &= E(Y_1^1) && \text{by Assumption 5} \\ &= E(Y_1 \mid A_1 = 1) && \text{by Assumptions 1, 2, 3} \\ E(Y_1^{0,0}) &= E(Y_1 \mid A_1 = 0, L = 0) && \text{by Assumptions 1, 2, 3, 4} \\ &\text{identified from } f_{Y_1|A_1=0}(y) && \text{by Assumption 6} \end{aligned}$$

Then,

$$\begin{aligned} E(Y_1^{a_1=1} - Y_0) - E(Y_1^{a_1=0,l=0} - Y_0) &= \big[E(Y_1^{a_1=1}) - E(Y_0)\big] - \Big[E(Y_1^{a_1=0,l=0}) - E(Y_0)\Big] \\ &= E(Y_1^{a_1=1}) - E(Y_1^{a_1=0,l=0}) \\ &= \Delta_{NR}. \end{aligned} \tag{17}$$

Note that $E(Y_1^{a_1=1}) - E(Y_1^{a_1=0,l=0}) = \Delta_{NR}$.

Result 3 (Average Treatment Effect for Placebo Responders). The average treatment effect for true placebo responders is

$$\Delta_{\text{PR}} = E(Y_1^{a_1=1,l=1}) - E(Y_1^{a_1=0,l=1}).$$

We can also summarize the ATE for placebo responders as the quantity $\Delta_{PR} = \Delta_{NR} - \Delta_{placebo}$.

Proof.

$$
\begin{aligned}
&E(Y_1^{a_1=1,l=1}) = E(Y_1^{a_1=1,l=0}) = E(Y_1 \mid A_1 = 1) && \text{by Assumptions } 5,1,2,3\\
&E(Y_1^{a_1=0,l=1}) = E(Y_1 \mid A_1 = 0, L = 1) && \text{by Assumptions } 1,2,3,4\\
&\qquad\text{identified from } f_{Y_1|A_1=0}(y) && \text{by Assumption } 6
\end{aligned}
$$

$$
\begin{aligned}
\Delta_{PR} &= E(Y_1^{a_1=1}) - E(Y_1^{a_1=0,l=1})\\
&= E(Y_1^{a_1=1}) - \left[E(Y_1^{a_1=0,l=0}) + \Delta_{\text{placebo}}\right]\\
\Rightarrow \Delta_{\text{PR}} &= \Delta_{\text{NR}} - \Delta_{\text{placebo}} \qquad \text{by algebra under Assumption 5.} \qquad (18)
\end{aligned}
$$

Result 4 (Average Treatment Effect for All Participants - Difference). The average treatment effect for all participants can be represented as a simple difference

$$\Delta_{\text{all}} = E(Y_1^{a_1=1}) - E(Y_1^{a_1=0}).$$

Proof.

$$
\begin{aligned}
&E(Y_1^a) = E(Y_1 \mid A_1 = a) && \text{by Assumptions } 1,2,3\\
&\Rightarrow \Delta_{\text{all}} = E(Y_1 \mid A_1 = 1) - E(Y_1 \mid A_1 = 0).
\end{aligned}
$$

Result 5 (Average Treatment Effect for All Participants - Mixture). The ATE for all participants can also be represented as a mixture

$$\Delta_{\text{all}} \;=\; (1 - P(L=1))\Delta_{\text{NR}} + P(L=1)\Delta_{\text{R}} \;=\; \Delta_{\text{NR}} - P(L=1)\Delta_{\text{placebo}}.$$

Proof. Next, the ATE for the full population is presented in mixture form. $P(L=1)$ is identified from $f_{Y_1|A_1=0}(y)$ by Assumptions 3 and 6. $\Delta_{\text{NR}},\ \Delta_{\text{PR}},\ \Delta_{\text{placebo}}$ are identified as in Results 1-3. Then from algebra,

$$
\begin{aligned}
E(Y_1^{a_1=1}) - E(Y_1^{a_1=0,l}) &= E(Y_1^{a_1=1}) - E\left(P(L=1)Y_1^{a_1=0,l=1} + (1-P(L=1))Y_1^{a_1=0,l=0}\right)\\
&= E(Y_1^{a_1=1}) - \left[P(L=1)E(Y_1^{a_1=0,l=1}) + (1-P(L=1))E(Y_1^{a_1=0,l=0})\right]\\
&= E(Y_1^{a_1=1}) - P(L=1)(\Delta_{\text{placebo}} + E(Y_1^{a_1=0,l=0})) - (1-P(L=1))E(Y_1^{a_1=0,l=0})\\
&= E(Y_1^{a_1=1}) - E(Y_1^{a_1=0,l=0}) - P(L=1)\Delta_{\text{placebo}}\\
&= \Delta_{NR} - P(L=1)\Delta_{\text{placebo}}. \qquad (19)
\end{aligned}
$$

Estimator Derivations

Result 6 (Expectation of Stage 1 Estimator).

$$E(\hat{\theta}_1) = \Delta_{\text{all}}.$$

Proof. We make Assumptions 1-7. Then,

$$\begin{aligned}
E(\hat{\theta}_1) &= E\left[\frac{1}{n_A}\sum_{i=1}^{n}(Y_{i,1}-Y_{i,0})I(A_{i,1}=1) - \frac{1}{n_P}\sum_{i=1}^{n}(Y_{i,1}-Y_{i,0})I(A_{i,1}=0)\right] \\
&= \frac{1}{n_A}\sum_{i=1}^{n}E(Y_{i,1}-Y_{i,0}|A_{i,1}=1) - \frac{1}{n_P}\sum_{i=1}^{n}E(Y_{i,1}-Y_{i,0}|A_{i,1}=0) \\
&= \frac{1}{n_A}\sum_{i=1}^{n_A}E(Y_{i,1}^{a_1=1}-Y_{i,0}) - \frac{1}{n_P}\sum_{i=1}^{n_P}E(Y_{i,1}^{a_1=0,l}-Y_{i,0}) \\
&= E(Y_1^{a_1=1}-Y_0) - E(Y_1^{a_1=0,l}-Y_0) \\
&= E(Y_1^{a_1=1}) - E(Y_1^{a_1=0,l}) \\
&= \Delta_{\text{all}}.
\end{aligned} \tag{20}$$

Result 7 (Expectation of the Stage 2 Estimator).

$$E(\hat{\theta}_2) = \Delta_{NR} - q_1\Delta_{\text{placebo}}.$$

$E(\hat{\theta}_2) = \Delta_{NR}$ if $q_1 = P(L=1|R=0) = 0$, i.e. if no true placebo responders $L=1$ are classified as placebo non-responders $R=0$.

Proof. Let $P(L=0,R=0) = p_{00}$ and $P(L=1,R=0) = p_{10}$ so that $q_l = P(L=l|R=0) = p_{l0}/(p_{00}+p_{10})$. To calculate these probabilities, we would need to make a distributional assumption about Y_t.

We also define

$$\begin{aligned}
E(Y_2^{a_1,a_2,l}|R=0) &= P(L=0|R=0)E(Y_2^{a_1,a_2,l=0}|R=0) \\
&\quad + P(L=1|R=0)E(Y_2^{a_1,a_2,l=1}|R=0) \\
&= q_0E(Y_2^{a_1,a_2,l=0}|R=0) + q_1E(Y_2^{a_1,a_2,l=1}|R=0)
\end{aligned} \tag{21}$$

Again we make Assumptions 1-7 (modified accordingly for Stage 2), so that the placebo non-responder average treatment effect is the same as from Stage 1, i.e.

$$E(Y_2^{a_1=0,a_2=1,l=0}) - E(Y_2^{a_1=0,a_2=0,l=0}) = E(Y_1^{a_1=1}) - E(Y_1^{a_0=0,l=0}) = \Delta_{NR}.$$

Note that given R, the sample sizes for each stage 2 arm are fixed n_{PA} and n_{PP} since the classified responder status determines the samples size of each arm. Then we get

$$
\begin{aligned}
E(\hat{\theta}_2) =& E\left[\frac{1}{n_{PA}}\sum_{i=1}^{n}(Y_{i,2}-Y_{i,1})I(A_{i,1}=0,R_i=0,A_{i,2}=1)|R\right] \\
&- E\left[\frac{1}{n_{PP}}\sum_{i=1}^{n}(Y_{i,2}-Y_{i,1})I(A_{i,1}=0,R_i=0,A_{i,2}=0)|R\right] \\
=& \frac{1}{n_{PA}}\sum_{i=1}^{n_{PA}} E\left[Y_{i,2}-Y_{i,1}|A_{i,1}=0,R_i=0,A_{i,2}=1,\right] \\
&- \frac{1}{n_{PP}}\sum_{i=1}^{n_{PP}} E\left[Y_{i,2}-Y_{i,1}|A_{i,1}=0,R_i=0,A_{i,2}=0\right] \\
=& \frac{n_{PA}}{n_{PA}}E\left[Y_2^{a_1=0,a_2=1,l}-Y_1^{a_1=0,l}|R=0\right] \\
&- \frac{n_{PP}}{n_{PP}}E\left[Y_2^{a_1=0,a_2=0,l}-Y_1^{a_1=0,l}|R=0\right] \\
=& E\left[Y_2^{a_1=0,a_2=1,l}-Y_1^{a_1=0,l}|R=0\right] \\
&- E\left[Y_2^{a_1=0,a_2=0,l}-Y_1^{a_1=0,l}|R=0\right] \\
=& E\left[Y_2^{a_1=0,a_2=1,l}|R=0\right]-E\left[Y_2^{a_1=0,a_2=0,l}|R=0\right] \\
=& E\left[Y_2^{a_1=0,a_2=1,l}|R=0\right]-E\left[Y_2^{a_1=0,a_2=0,l}|R=0\right] \\
=& \left[q_0E(Y_2^{a_1=0,a_2=1,l=0}|R=0)+q_1E(Y_2^{a_1=0,a_2=1,l=1}|R=0)\right] \\
&- \left[q_0E(Y_2^{a_1=0,a_2=0,l=0}|R=0)+q_1E(Y_2^{a_1=0,a_2=0,l=1}|R=0)\right] \\
=& q_0\left[E(Y_2^{a_1=0,a_2=1,l=0}|R=0)-E(Y_2^{a_1=0,a_2=0,l=0}|R=0)\right] \\
&+ q_1\left[E(Y_2^{a_1=0,a_2=1,l=1}|R=0)-E(Y_2^{a_1=0,a_2=0,l=1}|R=0)\right] \\
=& q_0\Delta_{NR}+q_1\Delta_{PR}, \quad \text{where } \Delta_{PR}=\Delta_{NR}-\Delta_{\text{placebo}} \text{ is the placebo response.} \\
=& q_0\Delta_{NR}+q_1(\Delta_{NR}-\Delta_{\text{placebo}}) \\
=& \Delta_{NR}(q_0+q_1)-q_1\Delta_{\text{placebo}} \\
\Rightarrow E(\hat{\theta}_2) =& \Delta_{NR}-q_1\Delta_{\text{placebo}} \neq \Delta_{\text{all}}.
\end{aligned} \tag{22}
$$

Result 8 (Expectation of the Weighted SPCD Estimator).

$$E(\hat{\theta}_w) = \Delta_{NR} - \Delta_{\text{placebo}}(wP(L=1)+(1-w)P(L=1|R=0)).$$

Proof.

$$
\begin{aligned}
E(\hat{\theta}_w) &= wE(\hat{\theta}_{1,\text{all}})+(1-w)E(\hat{\theta}_2) \\
&= w\Delta_{\text{all}}+(1-w)\left[\Delta_{NR}-q_1\Delta_{\text{placebo}}\right] \\
&= w(\Delta_{NR}-p_L\Delta_{\text{placebo}})+(1-w)(\Delta_{NR}-q_1\Delta_{\text{placebo}}) \\
&= \Delta_{NR}(w+(1-w))-\Delta_{\text{placebo}}(wp_L+(1-w)q_1) \\
&= \Delta_{NR}-\Delta_{\text{placebo}}(wP(L=1)+(1-w)P(L=1|R=0)).
\end{aligned} \tag{23}
$$

Result 9 (Weighted SPCD Estimator Targets Δ_{all}). $\hat{\theta}_w$ is unbiased for Δ_{all} if any of the following hold:

1. $P(L=1)=0$.

2. $\Delta_{\text{placebo}} = 0$.
3. $R \perp\!\!\!\perp L$.
4. We set $w = 1$.

Proof. [font=*]

Suppose $P(L = 1) = 0$.

By Result 5, $\Delta_{\text{all}} = \Delta_{NR}$. *Then,*

$$\begin{aligned} E(\hat{\theta}_w) &= \Delta_{NR} - \Delta_{\text{placebo}}(wP(L=1) + (1-w)P(L=1|R=0)) \\ &= \Delta_{NR} - \Delta_{\text{placebo}}(w0 + (1-w)0) \quad \textit{since } P(L=1) = 0 \Rightarrow P(L=1|R=0) = 0 \\ &= \Delta_{NR} = \Delta_{\text{all}}. \end{aligned} \tag{24}$$

2. *Suppose* $\Delta_{\text{placebo}} = 0$.

By Result 5, $\Delta_{\text{all}} = \Delta_{NR}$. *Then,*

$$\begin{aligned} E(\hat{\theta}_w) &= \Delta_{NR} - 0(wP(L=1) + (1-w)P(L=1|R=0)) \\ &= \Delta_{NR} = \Delta_{\text{all}}. \end{aligned} \tag{25}$$

3. Suppose $R \perp\!\!\!\perp L$.

Then $P(L = 1|R = 0) = P(L = 1) = p_L$, *and,*

$$\begin{aligned} E(\hat{\theta}_w) &= \Delta_{NR} - \Delta_{\text{placebo}}(wP(L=1) + (1-w)P(L=1|R=0)) \\ &= \Delta_{NR} - \Delta_{\text{placebo}}(wp_L + (1-w)p_L) \\ &= \Delta_{NR} - p_L\Delta_{\text{placebo}} \\ &= \Delta_{\text{all}}. \end{aligned} \tag{26}$$

Note $R \perp\!\!\!\perp L$ *is trivially true if* C_R *assigns every participant* $R_i = 0$.

4. Suppose we set $w = 1$.

Then $\hat{\theta}_w = \hat{\theta}_1$. *See Result 6 for the proof* $\hat{\theta}_1$ *is unbiased for* Δ_{all}.

Result 10 (Weighted SPCD Estimator Targets Δ_{NR}). $\hat{\theta}_w$ is unbiased for Δ_{NR} if $w = 0 \Rightarrow \hat{\theta}_w = \hat{\theta}_2$ and $\hat{\theta}_2$ is unbiased for Δ_{NR}.

Proof. See Result 7 for proof of when $\hat{\theta}_2$ is unbiased for Δ_{NR}.