The roles of estimands and assumptions in causal inference: Comment on “Chasing shadows: how implausible assumptions skew our understanding of causal estimands”

Sizhu Lu[*] and Peng Ding[†]

Department of Statistics, University of California, Berkeley

1 Introduction

1.1 Overview

We congratulate the authors for their insightful discussion on the role of assumptions in causal inference, particularly in the context of principal stratification estimands in clinical trials. Their examination of the tension between scientific relevance and practical learnability is an important contribution to the field. While acknowledging that the move toward model-free estimands is a valuable step, the authors highlight the challenge of balancing scientific relevance with practical learnability and emphasize the importance of thoroughly evaluating assumptions when identifying causal effects. We agree with their focus on the plausibility of the assumptions and most of their critiques on Qu et al. (2020). However, we argue that principal stratification remains highly relevant for formulating causal effects on subgroups defined by the potential values of post-treatment variables. We respectfully disagree with their critique of principal stratification estimands and advocate for improvements in identification strategies and the use of sensitivity analyses. Additionally, we highlight key distinctions between principal stratification and mediation analysis.

[*] 367 Evans Hall, Berkeley, CA 94720 USA. Email: sizhu_lu@berkeley.edu

[†] 425 Evans Hall, Berkeley, CA 94720, USA. Email: pengdingpku@berkeley.edu

1.2 Distinction between causal estimands and causal assumptions

Causal estimands and causal assumptions are two components of causal inference. The choice of causal estimands should correspond to the scientific question of interest, while the choice of causal assumptions should be guided by our understanding of the data-generating processes. Some causal estimands are relatively straightforward to identify, even allowing for multiple identification strategies under different sets of assumptions. In such cases, research may focus on developing the most efficient estimators for those causal estimands. However, other causal estimands are more challenging to identify, particularly when our understanding of the data-generating process is limited. In these situations, point identification may not be possible, and we may only obtain bounds on the estimand (e.g., Zhang and Rubin, 2003; Cheng and Small, 2006; Yang and Small, 2016). In some cases, point identification is achieved by imposing additional assumptions, which may be introduced for analytical convenience rather than being fully justified by our understanding of the data-generating process. When relying on such assumptions, it is crucial to conduct sensitivity analyses to assess the robustness of the results and evaluate the potential impact of violations of the assumptions (Jiang et al., 2022; Mattei et al., 2025).

2 Brief review of principal stratification

2.1 Why we need principal stratification

We first review the principal stratification strategy (Frangakis and Rubin, 2002), the causal estimand of interest, and the identification of the estimand. We follow the same notation as in Vansteelandt and Van Lancker (2025). Let $T \in \{0,1\}$ denote the binary treatment assignment, with $T = 1$ if assigned to the treatment arm, and $T = 0$ if assigned to the control arm, X denote the pre-treatment covariates, and Y denote the outcome of interest. Denote $A \in \{0,1\}$ as the adherence indicator after the treatment initiation, with $A = 1$ if the patient adhered to the assigned treatment until the measurement of Y, and $A = 0$ if the patient did not adhere. Both A and Y are post-treatment variables, therefore, under the potential outcome framework, they have potential values $A(t)$ and $Y(t)$ for $t = 0, 1$. The observed adherence status and outcome are $A = TA(1) + (1 - T)A(0)$ and $Y = TY(1) + (1 - T)Y(0)$, respectively.

Define the average treatment effect of the treatment assignment T on Y as $\tau_Y = E\{Y(1) - Y(0)\}$, the intention-to-treat estimand. In clinical trials, the treatment T is randomly assigned to patients with a known mechanism, therefore, τ_Y is identified as $\tau_Y = E(Y \mid T = 1) - E(Y \mid T = 0)$ by design if Y is measured for all patients, regardless of their adherence status. It is also known as the treatment policy estimand proposed in ICH E9 (R1) (2019). However, in practice, there are at least two main challenges. First, the intention-to-treat estimand may not ideally answer the clinically relevant questions and have poor generalizability when the proportion of adherence is low, as well as when the proportion of adherence changes dramatically in a future trial. Second, we may not be able to measure the outcome of interest for patients who did not adhere to the treatment assigned, for example, if they discontinued the study and dropped out.

To address these challenges, researchers have employed principal stratification and transferred the goalpost to the principal causal effects, as suggested in ICH E9 (R1) (2019). Targeting the principal stratification estimand, we classify the population of all patients into four strata according to the joint value of $\{A(1), A(0)\}$. For a binary A, there are four principal strata. In this discussion, we focus on the following principal causal effect:

$$\tau_{11} = E\{Y(1) - Y(0) \mid A(1) = 1, A(0) = 1\}, \tag{1}$$

which represents the average causal effect of T on Y, within the subgroup of patients who always adhere to the treatment assigned no matter which treatment arm they belong to.

If all non-adherence occurs immediately at treatment assignment, the problem reduces to a non-compliance problem, where instrumental variable methods can be applied (Angrist et al., 1996). However, in clinical trials, non-adherence is more complex, as it can occur at any time point between treatment assignment and outcome measurement. When non-adherence happens long after treatment assignment, it is hard to believe in the exclusion restriction assumption, which requires that the treatment assignment only affects outcome interest through adherence status. Thus, what we address here is a more general form of non-adherence problem that is different from standard non-compliance, as discussed in parts of the causal inference literature.

2.2 Classic identification results under the monotonicity and principal ignorability assumptions

In this subsection, we present the classic identification assumptions and results commonly used in the principal stratification literature. We discuss their implications, justifications, and possible limitations. To identify τ_{11} using observed data, we can invoke the following assumptions.

Assumption 1 *We impose the following assumptions:*

(a) Randomization and overlap: T is conditionally independent of all potential values of the post-treatment variables given X, and $0 < \mathrm{pr}(T = 1 \mid X) < 1$ with probability 1.

(b) Monotonicity: $A(1) \geq A(0)$.

(c) Principal ignorability: $Y(1) \perp\!\!\!\perp A(0) \mid A(1), X$ and $Y(0) \perp\!\!\!\perp A(1) \mid A(0), X$.

Assumption 1(a) is satisfied by design in clinical trials. The commonly used monotonicity Assumption 1(b) helps with the identification of the principal strata (Jiang et al., 2022). Here Assumption 1(b) imposes monotonicity on $A(t)$, the potential adherence status under treatment t, which is different from the standard monotonicity assumption applied to the treatment received indicator (Angrist et al., 1996). In the context of clinical trials, Assumption 1(b) states that adherence under the control arm is no greater than under treatment. It may be reasonable if the treatment is perceived as beneficial, since individuals in the treatment arm may be more likely to adhere upon experiencing better health improvements. However, if the treatment has side effects or leads to adverse events, the opposite direction of monotonicity, $A(1) \leq A(0)$, may be more reasonable, as adherence could be easier under the control condition. In practice, adherence patterns may even reflect a mixture of both. Whether and which monotonicity is reasonable depends on the specific context, and we do not attempt to justify either universally. Instead, our primary objective in introducing it is to achieve the identification of the joint distribution of $A(1)$ and $A(0)$ given the covariates X.

Assumption 1(c) requires that the conditional distribution of $Y(1)$ are the same between two principal strata $\{A(1) = 1, A(0) = 1\}$ and $\{A(1) = 1, A(0) = 0\}$ given the observed pre-treatment variables (Jo and Stuart, 2009; Ding and Lu, 2017; Feller et al., 2017). In general, it is strong and untestable. We recommend that researchers check what the assumption means even under a simple linear structural model. For example, if $Y(t)$ follows linear models $E\{Y(t) \mid A(1), A(0), X\} =$

$\beta_{t0} + \beta_{t1}A(t) + \beta_{t3}^{\mathrm{T}}X$ without the $A(1-t)$ term for $t = 0, 1$, the principal ignorability assumption holds (Jiang and Ding, 2021).

Under Assumption 1, τ_{11} is nonparametrically identified with various identification formulas (Ding and Lu, 2017; Jiang et al., 2022):

$$\begin{aligned}\tau_{11} &= E\left[\frac{p_0(X)}{p_0}\{E(Y \mid A=1, T=1, X) - E(Y \mid A=1, T=0, X)\}\right] \\ &= E\left\{\frac{p_0(X)}{p_0}\frac{A}{p_1(X)}\frac{T}{\pi(X)}Y\right\} - E\left\{\frac{A}{p_0}\frac{1-T}{1-\pi(X)}Y\right\} \\ &= E\left[\frac{A(1-T)/\{1-\pi(X)\}}{p_0}\{E(Y \mid A=1, T=1, X) - E(Y \mid T=0, A=1, X)\}\right],\end{aligned}$$

where $\pi(X) = \mathrm{pr}(T = 1 \mid X)$, $p_t(X) = \mathrm{pr}(A = 1 \mid T = t, X)$, and $p_t = \mathrm{pr}(A = 1 \mid T = t)$ for $t = 0, 1$.

2.3 Identification in Qu et al. (2020)

Focusing on a similar set of causal estimands, Qu et al. (2020) utilized additional information in observed post-treatment variables, denoted as Z, with potential values $Z(t)$ for $t = 0, 1$, and observed value $Z = TZ(1)+(1-T)Z(0)$. They imposed another set of conditional independence assumptions:

Assumption 2 *Impose the following conditional independence assumptions:*

(a) $A(t) \perp\!\!\!\perp \{Y(1), Y(0), Z(1-t)\} \mid Z(t), X$*, for* $t = 0, 1$.

(b) $Y(t) \perp\!\!\!\perp Z(1-t) \mid Z(t), X$ *for* $t = 0, 1$.

(c) $Z(1) \perp\!\!\!\perp Z(0) \mid X$.

Under Assumption 1(a) and Assumption 2, Qu et al. (2020) provided an identification formula for τ_{11}, as well as several other estimands. Assumption 2(b) is the principal ignorability assumption imposed on $Z(t)$ instead of $A(t)$. By imposing Assumption 2(a) and (c), Qu et al. (2020) avoided the monotonicity assumption on A, but introduced interpretations that might be unrealistic in practical applications, as discussed in Vansteelandt and Van Lancker (2025). Assumptions 2(b) and (c) together imply that $Y(t) \perp\!\!\!\perp Z(1-t) \mid X$ for $t = 0, 1$, which also seems too strong. We will revisit Qu et al. (2020)'s setting in Section 5.

3 What we agree with Vansteelandt and Van Lancker (2025) on

The authors highlight the importance of plausible assumptions for principal stratum estimands and illustrate why Assumptions 2(a) and (c) are unreasonable in some contexts. They stress the need to go beyond merely stating identification assumptions by providing thorough interpretation and critical assessment within the specific real-world application.

We fully agree that causal identification assumptions should be carefully evaluated within specific applications, especially in medication applications where both the sign and the magnitude of the treatment effect matter. Additionally, we acknowledge that identification assumptions in the principal stratification setup, such as Assumptions 1 and 2, are typically strong and untestable, which raises concerns about their practical plausibility. Thus, it is essential to be careful and transparent when we consider principal stratification estimands.

That said, in many cases, the principal stratum-specified causal effect is the most relevant parameter of interest, and its identification inherently requires strong assumptions. Progress requires making assumptions while carefully assessing their plausibility on a case-by-case basis. We must be transparent about these assumptions, and avoiding them altogether seems a pessimistic view of causal inference.

Additionally, since many identification assumptions are strong and untestable, sensitivity analysis must be an essential component for evaluating the robustness of causal conclusions. When the assumptions cannot be directly tested with observed data, sensitivity analysis allows us to explore how violations might affect the results. For instance, we can employ the sensitivity analysis framework proposed in Jiang et al. (2022) in principal stratification, which takes care of possible violations of both Assumptions 1(b) and (c). In practice, downstream decision-making should consider not only point estimates and inference under assumptions but also a broad range of sensitivity analyses to ensure more informed and reliable conclusions.

4 What we disagree with Vansteelandt and Van Lancker (2025) on

4.1 The critique of principal stratification estimands as a whole

While we acknowledge the authors' concerns about principal stratification, we believe that their rejection of these estimands is too broad. The principal stratification estimand remains useful when used judiciously, as it provides clinically meaningful causal parameters (Frangakis and Rubin, 2002; Rubin, 2006; VanderWeele, 2011; Mealli and Mattei, 2012; Jiang and Ding, 2021; Ding, 2024; Lu et al., 2023). The challenge lies not in the principal stratification estimand itself but in the choice of identification assumptions. Thus, we should distinguish between the estimand itself and the identification assumptions used in specific applications.

Rather than discarding it entirely, researchers should improve identification strategies and incorporate sensitivity analyses. We propose that instead of rejecting the use of the estimand, efforts should go toward developing better diagnostic tools and making it more practical. That said, again, we are not arguing that we should not examine the causal assumptions in Qu et al. (2020) critically. Indeed, the causal assumptions for principal stratification are strong and untestable.

4.2 Mediation analysis versus principal stratification

In this subsection, we argue that the assumptions in mediation analysis are not necessarily more plausible than those in principal stratification, and the choice between these frameworks should be guided by the specific research question and data structure. We consider the general case where $A(t)$ is any intermediate variable throughout the subsection. The authors claim in Section 5.1 that

> the cross-world independence assumptions used in mediation analysis tend to have *greater* plausibility than those previously discussed.

However, we would like to point out that the principal stratification and mediation analysis have fundamentally different focuses (Ding, 2024, Chapter 28.3). Principal stratification examines heterogeneous treatment effects across subgroups defined by joint values of $\{A(1), A(0)\}$, whereas mediation analysis treats A as a mediator in the causal pathway from T to Y. For example, the causal diagram in Figure 1 of Vansteelandt and Van Lancker (2025) indeed imposes causal relationships among T, A, and Y, which may not exist considering the principal stratification estimand. In some applications, A may not lie on the causal pathway from T to Y and may even be measured after the outcome Y.

We argue that many real-world problems in clinical trials are better suited to the principal stratification estimand, which identifies causal effects on subgroups defined by the joint values of $\{A(1), A(0)\}$. For example, τ_{11} describes the treatment effect for a clinically meaningful subgroup. Therefore, we disagree that mediation analysis provides a more appropriate framework for this research question.

Moreover, mediation analysis involves a-priori counterfactual potential outcomes, whereas principal stratification does not. Although the principal ignorability Assumption 1(c) and the cross-world independence assumption in mediation analysis both involve potential outcomes across different treatment conditions, the cross-world counterfactual independence assumption in mediation analysis involves nested potential outcomes such as $Y(t, A(1-t))$, which do not correspond to any hypothetical experiments. Therefore, we disagree with the claim that cross-world assumptions in mediation analysis are less restrictive than principal ignorability (Baccini et al., 2017; Forastiere et al., 2018). While these assumptions are not mutually nested, we find the cross-world independence assumption in mediation analysis conceptually less intuitive. See Andrews and Didelez (2021) for a more comprehensive review of the cross-world independence assumption.

That said, we do not dismiss the value of mediation analysis, which has important applications (VanderWeele, 2015). However, we believe that naively ranking the two frameworks in general can be misleading, as their suitability depends on the specific research question.

4.3 Do not throw the baby out with the bathwater

In conclusion, while we acknowledge the challenges with principal stratification, especially in terms of strong identification assumptions, it remains a valuable tool for understanding clinically meaningful causal effects. Rather than discarding principal stratification, we should focus on improving identification strategies and enhancing the robustness of the methods. By refining assumptions and conducting sensitivity analyses, principal stratification can provide meaningful insights, especially in clinical applications. Therefore, the priority should be to address its challenges and strengthen its applicability in real-world settings.

5 Some further comments on Qu et al. (2020)

5.1 Qu et al. (2020) implicitly imposes the unnecessarily strong assumption: conditional independence between $A(1)$ and $A(0)$ given X

In this subsection, we critically examine the identification results provided in Qu et al. (2020) and argue that their proposed identification formulas implicitly rely on a strong assumption: the conditional independence between $A(1)$ and $A(0)$ given X, which is not justified by their stated Assumption 2.

Qu et al. (2020) claimed that τ_{11} (which is their S_{++}) is identified under Assumptions 2, while we argue the identification formulas provided in their paper are only valid if they further assume $A(1) \perp\!\!\!\perp A(0) \mid X$. In Section A.4 in the appendix, Qu et al. (2020) wrote

> The probability for a patient to be adherent to both treatments can be expressed in two ways:
>
> $$\begin{aligned} & \operatorname{pr}(A(0)=1, A(1)=1) \\ = \;& E\left[\{\operatorname{pr}(T=1 \mid X)\}^{-1} I(T=1, A=1) \cdot E\{\operatorname{pr}(A(0)=1 \mid X, Z(0)) \mid X\}\right] \qquad (2) \\ = \;& E\left[\{\operatorname{pr}(T=0 \mid X)\}^{-1} I(T=0, A=1) \cdot E\{\operatorname{pr}(A(1)=1 \mid X, Z(1)) \mid X\}\right]. \end{aligned}$$

Taking (2) as an example, the left-hand side of (2) equals $E\left[E\{A(1)A(0) \mid X\}\right]$, while the right-hand side equals

$$\begin{aligned} E\left(\frac{TA}{\pi(X)} \underbrace{E\left[E\{A(0) \mid X, Z(0)\} | X\right]}_{=E\{A(0)|X\}}\right) &= E\left[\frac{E\{TA(1) \mid X\}}{\pi(X)} E\{A(0) \mid X\}\right] \\ &= E\left[E\{A(1) \mid X\} E\{A(0) \mid X\}\right]. \end{aligned}$$

Therefore, (2) holds if $A(1)$ and $A(0)$ are uncorrelated conditional on X. With binary A, it is equivalent to $A(1) \perp\!\!\!\perp A(0) \mid X$, which cannot be implied from the assumptions in Qu et al. (2020). Similar to the critiques in Section 3 of Vansteelandt and Van Lancker (2025), $A(1) \perp\!\!\!\perp A(0) \mid X$ is another strong and untestable independence assumption on the joint potential values.

Generally, identification for the joint distribution $\{A(1), A(0)\}$ is crucial for the identification of principal causal effects. Rather than relying on the unnecessarily strong conditional independence $A(1) \perp\!\!\!\perp A(0) \mid X$, we can consider alternative identification strategies. For instance, monotonicity provides a viable approach to achieve identification of the joint when A is binary. Additionally, copula methods offer a flexible way to model the joint distribution (Jiang and Ding, 2021), allowing researchers to incorporate domain knowledge in specifying reasonable copula functions and the correlation coefficient.

5.2 Reinterpreting the role of adherence indicator with augmented notation: a missing data perspective instead of principal stratification

In this subsection, we propose an augmented potential outcomes notation and discuss the role of the adherence indicator, exploring various interpretations and causal estimands. The augmented notation that depends jointly on treatment assignment t and adherence status a makes it more explicit what quantity is being targeted and what assumptions are needed for identification. This clarity is particularly useful when distinguishing between treatment policy effects, hypothetical full-adherence effects, and adherence as a mediating factor. Adopting this augmented notation helps disentangle these perspectives and encourages greater transparency in specifying both the estimand of interest and the role of the adherence indicator in the causal model.

Instead of the notation $Y(t)$ in Qu et al. (2020), we propose to augment the potential outcome as $Y(t,a)$, which depends not only on treatment assignment t but also on adherence status a. This notation explicitly accounts for the role of the adherence indicator. The observed outcome is then given by $Y = Y(T, A(T))$ under the *composition* assumption that $Y(t, A(t)) = Y(t)$ and the *consistency* assumption that $Y = Y(T)$. This alternative notation allows us to clarify different perspectives on the role of adherence indicator A.

First, under the treatment policy strategy, the target estimand is $E\{Y(1, A(1)) - Y(0, A(0))\}$, which is directly identifiable from the randomization of T in a randomized controlled trial.

Second, Qu et al. (2020) treated outcomes observed after non-adherence as confounded and censored, suggesting their primary interest lies in the potential outcome under full adherence. This leads to targeting the controlled direct effect $E\{Y(1,1) - Y(0,1)\}$, which corresponds to the hypothetical strategy proposed in ICH E9 (R1) (2019). Under our alternative formulation $Y(t,a)$, if $A(t) = 0$, the outcome of interest is unobserved, effectively reframing the problem as one of missing data rather than principal stratification. Here, the role of $A(t)$ is a missing indicator.

If we assume a missing at random mechanism, i.e., $A(t) \perp\!\!\!\perp \{Y(1,1), Y(0,1)\} \mid X$, then the inclusion of post-treatment variable $Z(t)$ does not contribute additional identifying information. However, in the application presented in Qu et al. (2020), the missing at random assumption appears unrealistic, which motivates their introduction of the post-treatment variable $Z(t)$, hoping that further conditioning on $Z(t)$ restores the missing at random property, i.e., $A(t) \perp\!\!\!\perp Y(t,1) \mid Z(t), X$ for $t = 0, 1$.

Under this assumption, the controlled direct effect $E\{Y(1,1)-Y(0,1)\}$ is identified by:

$$\begin{aligned}
E\{Y(1,1)\} &= E[E\{Y(1,1) \mid X\}] \\
&= E[E\{Y(1,1) \mid T=1, X\}] \\
&= E(E[E\{Y(1,1) \mid Z, T=1, X\} \mid T=1, X]) \\
&= E[E\{E(Y \mid A=1, Z, T=1, X) \mid T=1, X\}],
\end{aligned}$$

with a similar identification strategy for $E\{Y(0,1)\}$. It is the g-formula if we view (t,a) as the time-varying treatment under the standard sequential ignorability assumption (Hernán et al., 2000).

Third, adherence indicator A may itself lie on the causal pathway of T on Y, as illustrated in the causal diagram in Figure 1 of Vansteelandt and Van Lancker (2025). In this case, A is no longer merely a missing indicator, but functions as another treatment factor, warranting consideration of its causal effect on Y. For instance, $E\{Y(1,1)-Y(1,0)\}$ depicts the average treatment effect of adherence on outcome under treatment $T=1$.

In practice, for units assigned to the treatment arm, there may be multiple versions of non-adherence, each perhaps affecting outcomes in different ways. It challenges the stable treatment unit value assumption when using the notation $Y(t,a)$. Related issues on the existence of a wide variety of hypothetical scenarios are also discussed in ICH E9 (R1) (2019). More fundamentally, interventions on the adherence status can be ambiguous. A similar issue arises in the causal diagram presented in Figure 1 of Vansteelandt and Van Lancker (2025). By drawing an arrow from A to Y, the causal diagram suggests an implicit possibility of intervening on adherence status. This raises a series of questions: Should we allow for interventions on adherence status, and if so, how do we conceptualize such interventions? Given that adherence status may not always be directly manipulable, a more reasonable approach may be to consider stochastic intervention on adherence status. While treating adherence as a treatment factor is an intriguing direction, further development is needed.

5.3 Dealing with available data after non-adherence

In this subsection, we examine in detail how to appropriately handle available data after non-adherence when considering the principal stratification estimand, using the notation defined in the previous subsection.

For the principal stratum defined by $\{A(1) = 1, A(0) = 1\}$, we have

$$E\{Y(1, A(1)) - Y(0, A(0)) \mid A(1) = 1, A(0) = 1\} \quad = \quad E\{Y(1, 1) - Y(0, 1) \mid A(1) = 1, A(0) = 1\}.$$

Patients in this principal stratum would adhere to treatment under both arms. Therefore, it does not matter whether we use the treatment policy or the hypothetical strategy to deal with non-adherence because they would adhere under both treatment arms. However, for the other subgroups discussed in Qu et al. (2020), it is important to clearly state the strategy used to address non-adherence. For example, consider the subgroup defined by $A(1) = 1$, with no restrictions on $A(0)$ (denoted as S_{*+} in their paper). This group includes individuals who would adhere under treatment but may or may not adhere under control. In this case, the choice of estimand becomes crucial, as it determines how we should handle the available data after the non-adherence under control. When applying the treatment policy strategy, the target estimand is $E\{Y(1, A(1)) - Y(0, A(0)) \mid A(1) = 1\}$, which implies using all observed outcomes in the control group, regardless of the adherence status. It equals

$$\begin{aligned}
& E\{Y(1, A(1)) - Y(0, A(0)) \mid A(1) = 1\} \\
= \;& E\{Y(1, 1) \mid A(1) = 1\} - \sum_{a=0,1} E\{Y(0, a) \mid A(1) = 1, A(0) = a\}\text{pr}\{A(0) = a \mid A(1) = 1\} \\
= \;& E\{Y(1, 1) - Y(0, 0) \mid A(1) = 1, A(0) = 0\}\text{pr}\{A(0) = 0 \mid A(1) = 1\} \\
& + E\{Y(1, 1) - Y(0, 1) \mid A(1) = 1, A(0) = 1\}\text{pr}\{A(0) = 1 \mid A(1) = 1\},
\end{aligned}$$

which is a weighted average of two different causal effects within two different principal strata. In contrast, when applying the hypothetical strategy, the target estimand is $E\{Y(1, 1) - Y(0, 1) \mid A(1) = 1\}$, the identification of which requires imputing the counterfactual outcome $Y(0, 1)$ for patients in the control group who did not adhere. Overall, it is crucial to be explicit about both the strategy for handling non-adherence and the corresponding target estimand, as these choices can lead to substantially different scientific conclusions.

5.4 Defining principal strata based on adherence can lead to conceptual and practical issues

In this subsection, we further examine the principal causal effect estimand and argue that it may not be necessary to define principal strata based on adherence status. We also highlight that, even when we define principal strata as such, some estimands may pose conceptual challenges. We continue to use the notation $Y(t, a)$ from Section 5.2. Of course, for a post-treatment variable $A(t)$, principal stratification can still be applied to target estimands such as principal treatment effect $E\{Y(1,1) - Y(0,1) \mid A(1) = 1, A(0) = 1\}$ (Mattei et al., 2014a,b). However, we are critical on whether considering principal strata defined based on $\{A(1), A(0)\}$ is meaningful, particularly when they are purely random given $Z(t)$ and X, as required by Assumption 2(a). In such cases, it is more informative and practically relevant to consider principal strata defined based on the joint $\{Z(1), Z(0)\}$, which could provide a stronger connection to the heterogeneous treatment effect for various biological or behavioral values. Past literature provides rich results on principal stratification when Z is binary (Frangakis and Rubin, 2002; Ding and Lu, 2017; Jiang et al., 2022) and continuous (Lu et al., 2023; Zorzetto et al., 2024).

Finally, even if one chooses to define principal strata based on $\{A(1), A(0)\}$, as done in Qu et al. (2020), not all proposed estimands are conceptually reasonable. Among the four parameters in Qu et al. (2020), only τ_{11} is meaningful because $Y(t, 1)$ is only well-defined when $A(t) = 1$, and quantities like $Y(t, 1)$ for $A(t) = 0$ correspond to the nested potential outcomes that do not correspond to any hypothetical experiments. This is similar to the principal stratification in settings where the post-treatment variable is an indicator of death and the outcome is only well-defined for surviving patients (Zhang and Rubin, 2003; Rosenbaum, 2006; Ding et al., 2011). For example, consider the proposed estimand $E\{Y(1,1) - Y(0,1) \mid A(1) = 1\}$. The subgroup defined by $A(1) = 1$ is the combination of two principal strata: $\{A(1) = 1, A(0) = 1\}$ and $\{A(1) = 1, A(0) = 0\}$. Consequently, $E\{Y(1,1) - Y(0,1) \mid A(1) = 1\}$ is a weighted average of the treatment effects across the two principal strata. However, since $Y(t, 1)$ is only well-defined when $A(t) = 1$, attempting to identify or estimate $E\{Y(0,1) \mid A(1) = 1, A(0) = 0\}$ is not reasonable.

5.5 Did Qu et al. (2020) use principal stratification appropriately?

In summary, when A represents adherence status, whether we should view A as a missing indicator is questionable as discussed in Sections 5.2. Furthermore, even if we do treat A as a missing indicator, the information provided by τ_{11} is limited, which could lead to challenges in drawing broader causal conclusions, as discussed in Section 5.4. Overall, Qu et al. (2020) could have strengthened their analysis by more clearly stating their causal framework, specifically, by clarifying how they conceptualize the role of adherence and how this conceptualization informs the interpretation of principal strata defined by adherence status. Without a clear formulation of the causal estimands, proceeding with identification and estimation for the principal causal effects becomes problematic. This lack of clarity raises concerns about the appropriate use of the principal stratification estimand in their analysis.

6 When does additional post-treatment variable help in principal stratification?

In the classic setting of average treatment effect estimation, where the parameter of interest is $\tau_Y = E\{Y(1) - Y(0)\}$. If the observed post-treatment variables are predictive of the outcome, adjusting for those observed post-treatment variables might seem appealing. However, current literature has reached the consensus that such adjustments are problematic even when these variables are predictive of Y. A similar intuition applies in principal stratification. When the parameter of interest is the principal causal effect, $\tau_{11} = E\{Y(1)-Y(0) \mid A(1) = A(0) = 1\}$, and the principal strata are defined based on joint values $\{A(1), A(0)\}$, other post-treatment variables are unlikely to be helpful unless we make further strong and often untestable assumptions.

Next, consider the case when we target at the parameter $E\{Y(1) - Y(0) \mid A(1) = 1\}$, as in Qu et al. (2020). In their analysis, they imposed the principal ignorability assumption on Z but not on A. Principal ignorability for A assumes $A(t) \perp\!\!\!\perp Y(1-t) \mid A(1-t), X$. Consider the scenario where we suspect that $A(1-t)$ does not provide enough information to achieve independence, but we believe $Z(1-t)$ does. In this case, we impose the assumption $A(t) \perp\!\!\!\perp Y(1-t) \mid Z(1-t), X$. To identify $E\{Y(1) - Y(0) \mid A(1) = 1\} = E[E\{Y(1) - Y(0) \mid A(1) = 1, X\} \mid A(1) = 1]$, we need the identification for $E\{Y(0) \mid A(1) = 1, X\}$, $E\{Y(1) \mid A(1) = 1, X\}$, and the conditional distribution of X in the subgroup $A(1) = 1$. Identification for the latter two quantities follows directly from randomization. Consider $E\{Y(0) \mid A(1), X\}$, which can be written as

$$\begin{aligned} E\{Y(0) \mid A(1), X\} &= E[E\{Y(0) \mid A(1), Z(0), X\} \mid A(1), X] \\ &= E[E\{Y(0) \mid Z(0), X\} \mid A(1), X] \\ &= E(E[E\{Y(0) \mid Z(0), X\} \mid A(1), Z(1), X] \mid A(1), X) \\ &= E(E[E\{Y(0) \mid Z(0), X\} \mid Z(1), X] \mid A(1), X), \end{aligned}$$

where the last equality holds only if we further assume $A(t) \perp\!\!\!\perp Z(1-t) \mid Z(t), X$, as part of Assumption 2(a). Identifying $E[E\{Y(0) \mid Z(0), X\} \mid Z(1), X]$ requires identifying $Z(0) \mid Z(1), X$, which essentially requires identification of the joint distribution $\{Z(1), Z(0)\}$ given X. When Z is a one-dimensional binary variable, like in the standard principal stratification setup, the joint distribution of $Z(1)$ and $Z(0)$ is identified under randomization, provided we impose assumptions such as monotonicity. When Z is a multi-dimensional variable, the joint distribution of $\{Z(1), Z(0)\}$ is not identified even with the monotonicity assumption. This is likely the reason why Qu et al. (2020) imposed the conditional independence between $Z(0)$ and $Z(1)$ given X.

Thus, unless we have better domain knowledge to specify the joint distribution between $\{Z(1), Z(0)\}$ compared with $\{A(1), A(0)\}$, the role of the post-treatment variable Z is limited. However, if we are able to identify the joint $\{Z(1), Z(0)\}$, then $E\{Y(1) - Y(0) \mid A(1) = 1\}$ can be identified and $Z(t)$ would be indeed helpful.

Identification in principal stratification is often challenging without additional strong assumptions. A powerful strategy is to leverage auxiliary variables to improve the identification (Ding et al., 2011; Mattei and Mealli, 2011; Mealli and Pacini, 2013; Yang and Small, 2016; Yang and Ding, 2018; Jiang and Ding, 2021). Therefore, incorporating post-treatment variables to improve identification in principal stratification seems an appealing idea. Qu et al. (2020) made an interesting attempt to tackle this challenge. However, their attempt showed the fundamental difficulties associated with this approach. Additional post-treatment variables can only aid identification under strong and often untestable assumptions. Given these challenges, we remain skeptical about the applicability of such additional post-treatment variables, as they may not improve identification results without imposing additional strong assumptions.

That said, we are open to future research on leveraging additional post-treatment variables to improve principal stratification analysis, provided that additional assumptions are carefully justified by scientific knowledge in concrete applications.

Acknowledgement

We thank the Associate Editor for their constructive suggestions. We also thank Zhichao Jiang, Fan Li, Alessandra Mattei, Fabrizia Mealli, Fan Yang, and Junni Zhang for their helpful comments. Sizhu Lu and Peng Ding were supported by the U.S. National Science Foundation (grants # 1745640 and # 1945136).

References

Andrews, R. M. and Didelez, V. (2021). Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology*, 32(2):209–219.

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

Baccini, M., Mattei, A., and Mealli, F. (2017). Bayesian inference for causal mechanisms with application to a randomized study for postoperative pain control. *Biostatistics*, 18(4):605–617.

Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):815–836.

Ding, P. (2024). *A First Course in Causal Inference*. London: Chapman and Hall.

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

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

Feller, A., Mealli, F., and Miratrix, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics*, 42:726–758.

Forastiere, L., Mattei, A., and Ding, P. (2018). Principal ignorability in mediation analysis: through and beyond sequential ignorability. *Biometrika*, 105(4):979–986.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570.

ICH E9 (R1) (2019). International council for harmonisation of technical requirements for pharmaceuticals for human use: addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. pages 1–19.

Jiang, Z. and Ding, P. (2021). Identification of causal effects within principal strata using auxiliary variables. *Statistical Science*, 36(4):493–508.

Jiang, Z., Yang, S., and Ding, P. (2022). Multiply robust estimation of causal effects under principal ignorability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84:1423–1445.

Jo, B. and Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, 28:2857–2875.

Lu, S., Jiang, Z., and Ding, P. (2023). Principal stratification with continuous post-treatment variables: Nonparametric identification and semiparametric estimation. *arXiv preprint arXiv:2309.12425*.

Mattei, A., Ding, P., Ballerini, V., and Mealli, F. (2025). Assessing causal effects in the presence of treatment switching through principal stratification. *Bayesian Analysis*. in press.

Mattei, A. and Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:729–752.

Mattei, A., Mealli, F., and Pacini, B. (2014a). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics*, 70:278–288.

Mattei, A., Mealli, F., and Pacini, B. (2014b). Identification of local causal effects with missing outcome values and an instrument for non response. *Communications in Statistics-Theory and Methods*, 43(4):815–825.

Mealli, F. and Mattei, A. (2012). A refreshing account of principal stratification. *The International Journal of Biostatistics*, 8(1):1–19.

Mealli, F. and Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108:1120–1131.

Qu, Y., Fu, H., Luo, J., and Ruberg, S. J. (2020). A general framework for treatment effect estimators considering patient adherence. *Statistics in Biopharmaceutical Research*, 12(1):1–18.

Rosenbaum, P. R. (2006). Comment: the place of death in the quality of life. *Statistical Science*, 21(3):313–316.

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.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J. (2011). Principal stratification–uses and limitations. *The International Journal of Biostatistics*, 7(1):1–14.

Vansteelandt, S. and Van Lancker, K. (2025). Chasing shadows: How implausible assumptions skew our understanding of causal estimands. *Statistics in Biopharmaceutical Research*, this issue.

Yang, F. and Ding, P. (2018). Using survival information in truncation by death problems without the monotonicity assumption. *Biometrics*, 74(4):1232–1239.

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

Zhang, J. L. and 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.

Zorzetto, D., Canale, A., Mealli, F., Dominici, F., and Bargagli-Stoffi, F. J. (2024). Bayesian nonparametrics for principal stratification with continuous post-treatment variables. *arXiv preprint arXiv:2405.17669*.