Principal stratification analysis using principal scores
Author(s): Peng Ding and Jiannan Lu
Source: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 79, No. 3 (JUNE 2017), pp. 757-777
Published by: Oxford University Press for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/44681810
Accessed: 31-12-2023 08:20 +00:00

J. R. Statist. Soc. B (2017)
79, *Part* 3, *pp.* 757–777

Principal stratification analysis using principal scores

Peng Ding

University of California at Berkeley, USA

and Jiannan Lu

Microsoft Corporation, Redmond, USA

[Received May 2015. Final revision May 2016]

Summary. Practitioners are interested in not only the average causal effect of a treatment on the outcome but also the underlying causal mechanism in the presence of an intermediate variable between the treatment and outcome. However, in many cases we cannot randomize the intermediate variable, resulting in sample selection problems even in randomized experiments. Therefore, we view randomized experiments with intermediate variables as semiobservational studies. In parallel with the analysis of observational studies, we provide a theoretical foundation for conducting objective causal inference with an intermediate variable under the principal stratification framework, with principal strata defined as the joint potential values of the intermediate variable. Our strategy constructs weighted samples based on principal scores, defined as the conditional probabilities of the latent principal strata given covariates, without access to any outcome data. This principal stratification analysis yields robust causal inference without relying on any model assumptions on the outcome distributions. We also propose approaches to conducting sensitivity analysis for violations of the ignorability and monotonicity assumptions: the very crucial but untestable identification assumptions in our theory. When the assumptions required by the classical instrumental variable analysis cannot be justified by background knowledge or cannot be made because of scientific questions of interest, our strategy serves as a useful alternative tool to deal with intermediate variables. We illustrate our methodologies by using two real data examples and find scientifically meaningful conclusions.

Keywords: Causal inference; Exclusion restriction; Intermediate variable; Monotonicity; Non-parametric identification; Principal ignorability; Sensitivity analysis

1. Causal inference with intermediate variables

When an intermediate variable between a treatment and outcome exists, often researchers are interested in not only the average causal effect of the treatment on the outcome but also the underlying causal mechanism in the presence of the intermediate variable. Naive analysis by conditioning on the observed value of the intermediate variable does not yield valid causal interpretations without imposing strong assumptions. Principal stratification (Frangakis and Rubin, 2002), which is defined as the joint potential values of the intermediate variable under both treatment and control, can be viewed as a pretreatment covariate unaffected by the treatment. Therefore, conditioning on principal stratification yields subgroup causal effects.

The subgroup causal effects classified by principal stratification have clear scientific meanings in various settings. For instance, when the intermediate variable is the actual treatment received,

Address for correspondence: Peng Ding, Department of Statistics, University of California at Berkeley, 425 Evans Hall, Berkeley, CA 94720, USA.
E-mail: pengdingpku@berkeley.edu

 1369–7412/17/79757

the principal stratification variable indicates the compliance status, and the classical instrumental variable estimator identifies the average causal effect for compliers (Angrist *et al.*, 1996). When the intermediate variable is the indicator for survival status, the only sensible subgroup causal effect on the outcome is the effect for survivors who would potentially survive under both treatment and control (Rubin, 2006). When the intermediate variable is a surrogate for the outcome, we want to predict the causal effect on the outcome by the causal effect on the surrogate. An ideal surrogate must satisfy the causal necessity that zero effect on the surrogate implies zero effect on the outcome (Frangakis and Rubin, 2002) and the causal sufficiency that positive effect on the surrogate implies positive effect on the outcome (Gilbert and Hudgens, 2008). Therefore, we can assess these requirements for an ideal surrogate by conducting a principal stratification analysis.

Principal stratification clarifies causal inference with intermediate variables, but it also results in inferential difficulties because of the missingness of the principal stratification variable and the consequential mixture distributions of the observed data. We can sharpen inference about causal effects within principal strata only if we impose some of the following structural or modelling assumptions:

(a) *monotonicity*, that the treatment has a non-negative effect on the intermediate variable for each unit (e.g. Angrist *et al.* (1996) and Gilbert and Hudgens (2008));
(b) *exclusion restriction* (ER), that zero effect on the intermediate variable implies zero effect on the outcome (e.g. Angrist *et al.* (1996));
(c) *normal outcome distributions* within principal strata (e.g. Zhang *et al.* (2009) and Frumento *et al.* (2012));
(d) *additional covariates or secondary outcomes* (Ding *et al.*, 2011; Mattei and Mealli, 2011; Mattei *et al.*, 2013; Mealli and Pacini, 2013; Yang and Small, 2016; Jiang *et al.*, 2016).

For instance, the classical instrumental variable analysis requires an ER (Angrist *et al.*, 1996), which may not be justified by background knowledge or cannot be assumed because of the scientific questions of interest. Without an ER, Zhang *et al.* (2009) and Frumento *et al.* (2012) assumed normal distribution outcome models within principal strata, and thus identifiability of the causal effects within principal strata is ensured by identifiability of the normal mixture model. Unfortunately, the results are sensitive to the parametric modelling assumption, and the unbounded likelihood function jeopardizes statistical inference even with correctly specified models (Ding *et al.*, 2011; Frumento *et al.*, 2016). Without these assumptions, however, large sample bounds of causal effects are often too wide to be informative (Zhang and Rubin, 2003; Cheng and Small, 2006). We shall review more applications and further highlight the inferential difficulty of principal stratification without an ER in Section 2.

Recognizing the salient feature that the intermediate is not randomized even though the treatment is randomized, we take an alternative perspective, viewing the problem as a semi-observational study. For objective causal inference, Rubin (2007, 2008) advocated designs of observational studies without access to the outcome data, which prevents data snooping by selecting favourable outcome models. In parallel with this classical wisdom of propensity scores in observational studies (Rosenbaum and Rubin, 1983a), we propose to conduct principal stratification analysis based on principal scores, which are defined as the conditional probabilities of the latent principal strata given a rich set of covariates that ensure certain ignorability assumptions. Previously, applied researchers (Follman, 2000; Hill *et al.*, 2002; Jo and Stuart, 2009; Jo *et al.*, 2011; Stuart and Jo, 2015) used principal scores to analyse data subject to one-sided non-compliance, and theoretical researchers (Joffe *et al.*, 2007) suggested the use of principal scores to identify general causal effects within principal strata. We advance the literature by

providing the theoretical foundation for using principal scores in the analysis of randomized experiments with intermediate variables. To be more specific, we give the assumptions for identification, extend previous literature to deal with general principal stratification problems beyond one-sided non-compliance and propose statistically efficient and numerically stable weighting estimators for causal effects. The theoretical results allow for a two-step inferential procedure: we first construct weighted samples without access to the outcome data, and we then obtain simple weighting estimators for causal effects within principal strata. The whole inferential procedure does not involve any model assumptions of the outcomes, leading to more objective causal inference.

Furthermore, the central role of principal scores relies on certain ignorability and monotonicity assumptions. In parallel with sensitivity analysis in observational studies (Rosenbaum and Rubin, 1983b; Rosenbaum, 2002), we propose approaches to conducting sensitivity analysis for violations of the ignorability assumptions. Previous literature has either dealt with binary outcomes (e.g. Sjölander *et al.* (2009) and Schwartz *et al.* (2012)) or relied on modelling assumptions on the outcomes (e.g. Gilbert *et al.* (2003)), whereas our strategy deals with general outcomes and relies on fewer modelling assumptions. Other than a few exceptions (Zhang *et al.*, 2009; Ding *et al.*, 2011; Frumento *et al.*, 2012), most principal stratification analyses assumed monotonicity which might be too restrictive for some applications. Our sensitivity analysis technique further removes the monotonicity assumption and assesses the effect of its violations. The ignorability and monotonicity assumptions, though crucial for identifying the causal effects of interest, cannot be validated by observed data. Therefore, we argue that principal stratification analyses should always come with sensitivity analysis for violations of these assumptions.

The paper proceeds as follows. Section 2 reviews the basic framework of principal stratification. Section 3 defines principal scores and provides sufficient conditions for identifying causal effects within principal strata. Section 4 highlights the balancing properties of principal scores. Section 5 discusses estimation strategies that are efficient, stable and easy to implement. Section 6 proposes approaches to conducting sensitivity analyses for the ignorability and monotonicity assumptions. We conduct simulation studies in Section 7, apply our methodologies to real data examples in Section 8 and conclude in Section 9. We provide proofs and technical details in the on-line supplementary material.

The data and programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Potential outcomes and principal stratification

Consider a randomized controlled experiment with N units. We collect pretreatment covariates $\mathbf{X}_i$ for each unit i before the experiment. Let Z_i be the treatment assignment for unit i, with $Z_i = 1$ for treatment and $Z_i = 0$ for control. We also collect the outcome of interest Y_i for unit i, which can be general (continuous, binary, time to event, etc.). In practice, we may also collect some intermediate variables between the treatment and outcome that are helpful in explaining the underlying causal mechanism and treatment effect heterogeneity. We shall first focus on the case with a binary intermediate variable S, because the binary case has the widest applications as illustrated by examples 1–3 in the latter part of this section. We shall also comment on general S later in Section 9.

We use the potential outcomes framework to define causal effects. Under the stable unit treatment value assumption (Rubin, 1980), there is only one version of the treatment, and there is no interference between units. The stable unit treatment value assumption allows us to

define the potential values of the intermediate variable and outcome for unit i as $S_i(t)$ and $Y_i(t)$ under treatment t for $t=0$ and $t=1$. Completely randomized experiments satisfy the following treatment assignment mechanism, which we shall make use of throughout the paper.

Assumption 1 (randomization). $Z \perp\!\!\!\perp \{S(1), S(0), Y(1), Y(0), \mathbf{X}\}$.

Frangakis and Rubin (2002) introduced the notion of principal stratification, which is defined as the joint potential values of the intermediate variable $U_i = \{S_i(1), S_i(0)\} \in \{0,1\}^2$. For simplicity, we relabel the possible values of U, $\{(1,1),(1,0),(0,1),(0,0)\}$, as $\{ss, s\bar{s}, \bar{s}s, \bar{s}\bar{s}\}$ respectively. Because the principal stratification variable is unaffected by the treatment, inference conditionally on U yields a subgroup causal interpretation, which is captured by the following principal causal effect (PCE):

$$\text{ACE}_u = E\{Y(1) - Y(0) | U = u\} \qquad (u = ss, s\bar{s}, \bar{s}s, \bar{s}\bar{s}).$$

The notion of a PCE is not only mathematically sound for causal evaluations, but also of scientific interest in practice. Below, we review some important empirical applications and discuss the scientific meanings of PCEs in each case.

2.1. Example 1 (non-compliance)

Let S denote the actual treatment received, and non-compliance occurs if the treatment assignment differs from the treatment received. Angrist *et al.* (1996) called ss, $s\bar{s}$, $\bar{s}s$ and $\bar{s}\bar{s}$ always-takers, compliers, defiers and never-takers respectively.

2.2. Example 2 (truncation by death)

When some units die before measurements of their outcomes Y, the truncation-by-death problem occurs. Let S be the survival status, with $S=1$ for survival and $S=0$ for dead. For dead patients with $S=0$, the corresponding outcome Y is not well defined. Rubin (2006) argued that the only scientifically meaningful subgroup causal effect is ACE_{ss}: the survivor average causal effect, which is defined as the average causal effect among units who will potentially survive under both treatment and control. Other subgroup causal effects are not well defined owing to their unmeasured outcome under either treatment or control or both.

There are at least two problems that are similar to truncation by death. In labour economics where S is employment status and Y is income, the only sensible causal effect is ACE_{ss}: the average causal effect among the always employed units (Zhang and Rubin, 2003; Zhang *et al.*, 2009). In vaccine trials where S is infection status and Y is a post-infection outcome, we are interested in the causal effect of vaccine on the post-infection outcome among units who would develop infection under both treatment and control (Gilbert *et al.*, 2003; Hudgens and Halloran, 2006).

2.3. Example 3 (surrogate)

Surrogates are of great importance in clinical trials, when the measurement of the primary outcome is costly or time consuming. Let S denote the surrogate for the outcome Y. The goal of using the surrogate is to predict the causal effect on the outcome by the causal effect on the surrogate. Frangakis and Rubin (2002) argued that a good surrogate should satisfy 'causal necessity', i.e., whenever the treatment has no effect on the surrogate, it has no effect on the outcome ($\text{ACE}_{ss}=0$ and $\text{ACE}_{\bar{s}\bar{s}}=0$). As a complement, Gilbert and Hudgens (2008) further argued that a good surrogate also should satisfy 'causal sufficiency', i.e., whenever the treatment affects the surrogate, it also affects the outcome ($\text{ACE}_{s\bar{s}} \neq 0$ and $\text{ACE}_{\bar{s}s} \neq 0$). no direct effect, must have indirect effect

In practice, a particular data set may simultaneously have more than one of the problems that were discussed in examples 1–3 (Mattei and Mealli, 2007; Frumento *et al.*, 2012). In all the examples above, estimation of ACE_u is crucial for the substantive questions of interest. However, the inherent missingness of U, due to the ability of measuring only one of $S(1)$ and $S(0)$, jeopardizes the identification of the PCEs without some additional assumptions. In what follows, we review some commonly used assumptions and discuss their plausibility and limitations.

Assumption 2 (strong monotonicity). $S_i(0) = 0$ for all i.

In example 1 of non-compliance, when the control units have no access to receive the active treatment, strong monotonicity holds by the design of experiments. It is sometimes referred to as one-sided non-compliance (Imbens and Rubin, 2015), which allows us to rule out the always-takers ($U = ss$) and defiers ($U = \bar{s}s$). In the literature on surrogates, strong monotonicity is closely related to the 'constant biomarker' assumption (Gilbert and Hudgens, 2008).

Assumption 3 (monotonicity). $S_i(1) \geqslant S_i(0)$ for all i.

Monotonicity rules out only the defiers ($U = \bar{s}s$). In general, we cannot test monotonicity by using the observed data unless $\Pr(S = 1|Z = 1) < \Pr(S = 1|Z = 0)$.

Assumption 4 (ER). $Y_i(1) = Y_i(0)$ for $U_i = ss$ and $U_i = \bar{s}\bar{s}$.

The ER implies that $ACE_{ss} = ACE_{\bar{s}\bar{s}} = 0$. In example 1 of non-compliance, an ER is plausible in double-blinded trials because the outcome may be affected only by the treatment received. Angrist *et al.* (1996) showed that, under monotonicity and an ER, the complier average causal effect $ACE_{s\bar{s}}$ is identified by the ratio of the average causal effects on Y and S.

However, in many open-label trials, the treatment assignment may have a 'direct effect' on the outcome, and an ER may not hold (e.g. Hirano *et al.* (2000)). What is more important, we cannot assume an ER in the truncation-by-death and surrogate problems, because in these settings it is the question of concern to test whether ACE_{ss} or $ACE_{\bar{s}\bar{s}}$ is 0. Imposing an ER immediately discards the very scientific question of interest, which is not reasonable. Unfortunately, if we do not impose an ER because of either background knowledge or substantive questions of interest, we can no longer non-parametrically identify the PCEs without further assumptions. In this paper, we shall discuss alternative sufficient conditions that ensure non-parametric identifiability of the PCEs, and we propose estimators that rely on minimal modelling assumptions.

3. Non-parametric identification of principal causal effects

Our identification strategy, in parallel with the notion of propensity scores in observational studies, exploits principal scores defined as the probabilities of the latent principal strata given a rich set of pretreatment covariates. Although in the existing literature principal scores are used in the one-sided non-compliance problem, its rigorous theoretical foundation is lacking, and more importantly it cannot deal with more general cases. We fill in the gap by demonstrating some general identification results based on principal scores.

3.1. Principal scores

Although we cannot uniquely recover the unobserved principal strata indicators, we can create weighted samples based on principal scores:

$$e_u(\mathbf{X}) = \Pr(U = u|\mathbf{X}) \qquad (u = ss, s\bar{s}, \bar{s}s, \bar{s}\bar{s}).$$

Define $p_1 = \Pr(S = 1|Z = 1)$ and $p_0 = \Pr(S = 1|Z = 0)$ as the probabilities of S under treatment

and control, and the analogous conditional probabilities as $p_1(\mathbf{X}) = \Pr(S = 1 | Z = 1, \mathbf{X})$ and $p_0(\mathbf{X}) = \Pr(S = 1 | Z = 0, \mathbf{X})$ given covariates $\mathbf{X}$.

Under strong monotonicity, two strata $s\bar{s}$ and $\bar{s}\bar{s}$ exist. The observed data with $(Z = 1, S = 1)$ contain only strata $U = s\bar{s}$, and the observed data with $(Z = 1, S = 0)$ contain only strata $U = \bar{s}\bar{s}$. Therefore, we can use the treatment arm to identify the principal scores by $e_{s\bar{s}}(\mathbf{X}) = p_1(\mathbf{X})$ and $e_{\bar{s}\bar{s}}(\mathbf{X}) = 1 - p_1(\mathbf{X})$, and the proportions of the two principal strata by $\pi_{s\bar{s}} = p_1$ and $\pi_{\bar{s}\bar{s}} = 1 - p_1$.

Under monotonicity, three strata $ss, s\bar{s}$ and $\bar{s}\bar{s}$ exist. The observed data with $(Z = 1, S = 0)$ contain only strata $U = \bar{s}\bar{s}$, and the observed data with $(Z = 0, S = 1)$ contain only strata $U = ss$. Therefore, we can identify the principal scores by $e_{ss}(\mathbf{X}) = p_0(\mathbf{X})$, $e_{\bar{s}\bar{s}}(\mathbf{X}) = 1 - p_1(\mathbf{X})$ and $e_{s\bar{s}}(\mathbf{X}) = p_1(\mathbf{X}) - p_0(\mathbf{X})$, and the proportions of the three principal strata by $\pi_{ss} = p_0$, $\pi_{\bar{s}\bar{s}} = 1 - p_1$ and $\pi_{s\bar{s}} = p_1 - p_0$.

The above discussion demonstrates non-parametric identification of principal scores under (strong) monotonicity. We postpone the discussion of modelling principal scores to Section 4.2.

3.2. Principal ignorability and non-parametric identification

The observed data are mixtures of at most two latent principal strata. Our goal is to disentangle the latent components of the outcome distributions. Although we can view them as the weights for the latent subgroup indicators, principal scores themselves alone are not sufficient to identify the PCEs. The following principal ignorability (PI) assumptions (Jo and Stuart, 2009; Jo *et al.*, 2011; Stuart and Jo, 2015), in parallel with the ignorability assumption in observational studies (Rosenbaum and Rubin, 1983a), are sufficient conditions to identify the PCEs non-parametrically.

3.2.1. Under strong monotonicity

We invoke the following version of PI.

Assumption 5 (PI). $Y(0) \perp\!\!\!\perp U | \mathbf{X}$.

A weaker version of assumption 5, which is shown in Section 6.1, also suffices for our later discussion on identification. Here we use a stronger version for easy interpretation. PI assumes conditional independence of $Y(0)$ and U given $\mathbf{X}$, i.e. a random allocation of the principal stratification variable with respect to the control potential outcome given $\mathbf{X}$. PI requires an adequate set of covariates $\mathbf{X}$, conditionally on which there is no difference between the distributions of the control potential outcomes across principal strata $U = s\bar{s}$ and $U = \bar{s}\bar{s}$.

With the identifiability of the principal scores, PI further helps to identify ACE_u.

Proposition 1. Under strong monotonicity and PI, we can identify the PCEs by

$$\text{ACE}_{s\bar{s}} = E(Y | Z = 1, S = 1) - E\{w_{s\bar{s}}(\mathbf{X}) Y | Z = 0\},$$
$$\text{ACE}_{\bar{s}\bar{s}} = E(Y | Z = 1, S = 0) - E\{w_{\bar{s}\bar{s}}(\mathbf{X}) Y | Z = 0\},$$

where $w_{s\bar{s}}(\mathbf{X}) = e_{s\bar{s}}(\mathbf{X})/\pi_{s\bar{s}}$ and $w_{\bar{s}\bar{s}}(\mathbf{X}) = e_{\bar{s}\bar{s}}(\mathbf{X})/\pi_{\bar{s}\bar{s}}$.

The treatment group does not involve mixture distributions. The control group is a mixture of two strata $s\bar{s}$ and $\bar{s}\bar{s}$, and proposition 1 shows that the weight $w_u(\mathbf{X})$ is the ratio of the principal score over the marginal proportion of stratum u.

3.2.2. Under monotonicity

We invoke the following general principal ignorability (GPI).

Assumption 6 (GPI). $Y(z) \perp\!\!\!\perp U|\mathbf{X}$ for $z=0$ and $z=1$.

Again, a weaker version of GPI suffices to identify PCEs as discussed in Section 6.1, but the stronger version enjoys easier interpretation. The mathematical form of assumption 6 is similar to the ignorability assumption in observational studies (Rosenbaum and Rubin, 1983a), with U being the latent principal stratification variable instead of the treatment indicator. Intuitively, the conditional independence of GPI requires enough covariates $\mathbf{X}$ to remove all 'confounding' between U and Y. More precisely, conditionally on $\mathbf{X}$, there is no difference between the distributions of the treatment potential outcomes across strata $U=ss$ and $U=s\bar{s}$, and no difference between the distributions of the control potential outcomes across strata $U=\bar{s}\bar{s}$ and $U=s\bar{s}$. These interpretations will become more apparent in Section 6.1. See Guo *et al.* (2014) for a slightly different view on GPI.

Proposition 2. Under monotonicity and GPI, we can identify the PCEs by

$$\begin{aligned}
\mathrm{ACE}_{s\bar{s}} &= E\{w_{1,s\bar{s}}(\mathbf{X})Y|Z=1,S=1\} - E\{w_{0,s\bar{s}}(\mathbf{X})Y|Z=0,S=0\},\\
\mathrm{ACE}_{\bar{s}\bar{s}} &= E(Y|Z=1,S=0) - E\{w_{0,\bar{s}\bar{s}}(\mathbf{X})Y|Z=0,S=0\},\\
\mathrm{ACE}_{ss} &= E\{w_{1,ss}(\mathbf{X})Y|Z=1,S=1\} - E(Y|Z=0,S=1),
\end{aligned}$$

where

$$\begin{aligned}
w_{1,s\bar{s}}(\mathbf{X}) &= \frac{e_{s\bar{s}}(\mathbf{X})}{e_{s\bar{s}}(\mathbf{X})+e_{ss}(\mathbf{X})} \Big/ \frac{\pi_{s\bar{s}}}{\pi_{s\bar{s}}+\pi_{ss}},\\
w_{0,s\bar{s}}(\mathbf{X}) &= \frac{e_{s\bar{s}}(\mathbf{X})}{e_{s\bar{s}}(\mathbf{X})+e_{\bar{s}\bar{s}}(\mathbf{X})} \Big/ \frac{\pi_{s\bar{s}}}{\pi_{s\bar{s}}+\pi_{\bar{s}\bar{s}}},\\
w_{0,\bar{s}\bar{s}}(\mathbf{X}) &= \frac{e_{\bar{s}\bar{s}}(\mathbf{X})}{e_{s\bar{s}}(\mathbf{X})+e_{\bar{s}\bar{s}}(\mathbf{X})} \Big/ \frac{\pi_{\bar{s}\bar{s}}}{\pi_{s\bar{s}}+\pi_{\bar{s}\bar{s}}},\\
w_{1,ss}(\mathbf{X}) &= \frac{e_{ss}(\mathbf{X})}{e_{s\bar{s}}(\mathbf{X})+e_{ss}(\mathbf{X})} \Big/ \frac{\pi_{ss}}{\pi_{s\bar{s}}+\pi_{ss}}.
\end{aligned}$$

The observed data with $(Z=1,S=0)$ and $(Z=0,S=1)$ do not involve mixture distributions. The observed data with $(Z=1,S=1)$ contain a mixture of two strata $s\bar{s}$ and ss, and the weight $w_{1,u}(\mathbf{X})$ is the probability of stratum u conditionally on $(Z=1,S=1,\mathbf{X})$ divided by the probability conditional only on $(Z=1,S=1)$. A similar discussion applies to the observed data with $(Z=0,S=0)$ and the weight $w_{0,u}(\mathbf{X})$.

4. Balancing properties of principal scores

4.1. Balancing properties

Principal scores play a crucial role in the theory that was developed in the previous section. Therefore, it is of practical importance to select a principal score model that is close to the truth. Fortunately, we can use the following balancing conditions for any function of the covariates, $h(\mathbf{X})$, to guide our choice of the model for $\Pr(U|\mathbf{X})$.

Corollary 1. Under strong monotonicity, we have

$$\begin{aligned}
E\{h(\mathbf{X})|Z=1,S=1\} &= E\{w_{s\bar{s}}(\mathbf{X})h(\mathbf{X})|Z=0\},\\
E\{h(\mathbf{X})|Z=1,S=0\} &= E\{w_{\bar{s}\bar{s}}(\mathbf{X})h(\mathbf{X})|Z=0\}.
\end{aligned}$$

Corollary 2. Under monotonicity, we have

$$E\{w_{1,s\bar{s}}(\mathbf{X})h(\mathbf{X})|Z=1,S=1\}=E\{w_{0,s\bar{s}}(\mathbf{X})h(\mathbf{X})|Z=0,S=0\},$$
$$E\{h(\mathbf{X})|Z=1,S=0\}=E\{w_{0,\bar{s}\bar{s}}(\mathbf{X})h(\mathbf{X})|Z=0,S=0\},$$
$$E\{w_{1,ss}(\mathbf{X})h(\mathbf{X})|Z=1,S=1\}=E\{h(\mathbf{X})|Z=0,S=1\}.$$

Corollaries 1 and 2 are direct applications of propositions 1 and 2. Intuitively, because any functions of the covariates $h(\mathbf{X})$ are unaffected by the treatment within principal strata, the 'PCEs' on $h(\mathbf{X})$ are all 0s. Although simple, the balancing conditions in corollaries 1 and 2 allow for model checking for principal scores and are therefore of practical importance. If the balancing conditions above are violated, we need to build a more flexible model to account for the residual dependence of U on $\mathbf{X}$. For example, we can add higher order polynomial and interaction terms of the covariates to the logistic model, until the balancing conditions are well satisfied. This idea is similar to designs of observational studies for achieving objective causal inference (Rubin, 2007, 2008). When constructing weighted samples, we do not have access to the outcome data, because we require only $(\mathbf{X}, Z, S)$ for creating the principal score estimates. This outcome-free strategy for designs, which was advocated by Rubin (2007, 2008) and Imbens and Rubin (2015), has the merit of being free of data snooping based on repeated search for favourable outcome models.

4.2. Estimating principal scores

Although we have non-parametric identification results under the PI assumptions 5 and 6, we can easily deal with only low dimensional and discrete covariates to estimate the principal scores. With high dimensional or continuous covariates, we need to specify models for $\Pr(U|\mathbf{X})$.

Under strong monotonicity, U takes only two values, and we can use a logistic model for $\Pr(U|\mathbf{X})$. By randomization, we can fit a logistic model of S on $\mathbf{X}$ using only the data from the treatment group, because, within arm $Z=1$, we have $S=1$ if and only if $U=s\bar{s}$, and $S=0$ if and only if $U=\bar{s}\bar{s}$.

Under monotonicity, U takes three values; we can model $\Pr(U|\mathbf{X})$ as a three-level multinomial logistic model and use the EM algorithm (Dempster *et al.*, 1977) to find the maximum likelihood estimates by treating U as missing data. See the on-line supplementary material for computational details.

In practice, a correct specification of the principal score model $\Pr(U|\mathbf{X})$ is crucial for the validity of the PCE estimation, because misspecification of $\Pr(U|\mathbf{X})$ may lead to biased estimators for the PCEs. After fitting a principal score model, we can use corollaries 1 and 2 to check balance of some important covariates and their functions. If the balancing conditions are violated, we can fit a more flexible model (e.g. adding high order polynomials or interaction terms of $\mathbf{X}$ to the logistic models) until the balance conditions are satisfied.

5. Modelling the outcome and model-assisted estimators

Previous identification and sensitivity analysis results assume infinite amounts of data or a known distribution of the observed data. In this section, we discuss finite sample estimators of PCEs. For simplicity, in the main text we shall discuss only the estimator for $\mathrm{ACE}_{s\bar{s}}$ under monotonicity. We have similar results for other strata, the cases under strong monotonicity, and the cases for sensitivity analysis; we relegate the technical details to the on-line supplementary material.

The identification formulae in propositions 1–6 immediately give us simple moment estimators by weighting, with $e_u(\mathbf{X})$ and π_u replaced by their consistent estimators, and the expectations over the population replaced by their sample analogues. In the above discussion about identification and moment estimators for PCEs, we use the covariates to predict latent strata and to create weights. In fact, covariates contain useful information about both the principal strata and the outcome distributions. Now we shall use covariate adjustment to improve statistical efficiency for estimation. Covariate adjustment is based on the following simple fact that, for all u and all fixed vectors $\beta_{z,u}$,

$$\mathrm{ACE}_u = E\{Y(1) - \beta_{1,u}^{\mathrm{T}}\mathbf{X}|U=u\} - E\{Y(0) - \beta_{0,u}^{\mathrm{T}}\mathbf{X}|U=u\} + (\beta_{1,u} - \beta_{0,u})^{\mathrm{T}} E(\mathbf{X}|U=u). \quad (1)$$

Treating the 'residual' $Y(z) - \beta_{z,u}^{\mathrm{T}}\mathbf{X}$ as a new 'potential outcome', we can apply proposition 2 to identify three expectation terms in formula (1) via the following corollary.

Corollary 3. Under monotonicity and GPI, we have

$$\begin{aligned} E\{Y(1) - \beta_{1,s\bar{s}}^{\mathrm{T}}\mathbf{X}|U=s\bar{s}\} &= E\{w_{1,s\bar{s}}(\mathbf{X})(Y - \beta_{1,s\bar{s}}^{\mathrm{T}}\mathbf{X})|Z=1, S=1\}, \\ E\{Y(0) - \beta_{0,s\bar{s}}^{\mathrm{T}}\mathbf{X}|U=s\bar{s}\} &= E\{w_{0,s\bar{s}}(\mathbf{X})(Y - \beta_{0,s\bar{s}}^{\mathrm{T}}\mathbf{X})|Z=0, S=0\}, \\ E(\mathbf{X}|U=s\bar{s}) &= E\{w_{1,s\bar{s}}(\mathbf{X})\mathbf{X}|Z=1, S=1\} \\ &= E\{w_{0,s\bar{s}}(\mathbf{X})\mathbf{X}|Z=0, S=0\}. \end{aligned}$$

Define $n_{zs} = \#\{i : Z_i = z, S_i = s\}$. The covariate-adjusted estimator for $\mathrm{ACE}_{s\bar{s}}$ is

$$\begin{aligned} \widehat{\mathrm{ACE}}_{s\bar{s}}^{\mathrm{adj}} = {} & \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \hat{w}_{1,s\bar{s}}(\mathbf{X}_i)(Y_i - \beta_{1,s\bar{s}}^{\mathrm{T}}\mathbf{X}_i) - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \hat{w}_{0,s\bar{s}}(\mathbf{X}_i)(Y_i - \beta_{0,s\bar{s}}^{\mathrm{T}}\mathbf{X}) \\ & + \frac{1}{n_{11}+n_{00}} (\beta_{1,s\bar{s}} - \beta_{0,s\bar{s}})^{\mathrm{T}} \left\{ \sum_{\{i:Z_i=1,S_i=1\}} \hat{w}_{1,s\bar{s}}(\mathbf{X}_i)\mathbf{X}_i + \sum_{\{i:Z_i=0,S_i=0\}} \hat{w}_{0,s\bar{s}}(\mathbf{X}_i)\mathbf{X}_i \right\}. \end{aligned} \quad (2)$$

As long as the potential outcomes are correlated with the covariates, the 'residual potential outcomes', $Y(z) - \beta_{z,u}^{\mathrm{T}}\mathbf{X}$, will have smaller variances than the original potential outcomes. Therefore, the covariate-adjusted estimator in formula (2) tends to have a smaller asymptotic variance than the unadjusted estimator. Our simulation studies have verified this intuition. Although corollary 3 and the covariate-adjusted estimator in formula (2) hold for any fixed vectors $\beta_{1,s\bar{s}}$ and $\beta_{0,s\bar{s}}$, we need to choose them in practice. Intuitively, we can choose $\beta_{z,s\bar{s}}$ as the linear regression coefficient of $Y(z)$ on the space spanned by $\mathbf{X}$ for units $U = s\bar{s}$, i.e.

$$\beta_{z,s\bar{s}} = E(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U=s\bar{s})^{-1} E\{\mathbf{X}Y(z)|U=s\bar{s}\}.$$

Similarly to corollary 3, each component of the above least squares formula is also identifiable.

Corollary 4. Under monotonicity and GPI, we have

$$\begin{aligned} E\{\mathbf{X}Y(1)|U=s\bar{s}\} &= E\{w_{1,s\bar{s}}(\mathbf{X})\mathbf{X}Y|Z=1, S=1\}, \\ E\{\mathbf{X}Y(0)|U=s\bar{s}\} &= E\{w_{0,s\bar{s}}(\mathbf{X})\mathbf{X}Y|Z=0, S=0\}, \\ E(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U=s\bar{s}) &= E\{w_{1,s\bar{s}}(\mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}}|Z=1, S=1\} \\ &= E\{w_{0,s\bar{s}}(\mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}}|Z=0, S=0\}. \end{aligned}$$

Therefore, we choose $\beta_{1,s\bar{s}}$ as the weighted least squares regression coefficient of Y_i on $\mathbf{X}_i$ by using samples with $(Z_i=1, S_i=1)$ and weights $w_{1,s\bar{s}}(\mathbf{X}_i)$, and $\beta_{0,s\bar{s}}$ as the weighted least squares regression coefficient of Y_i on $\mathbf{X}_i$ by using samples with $(Z_i=0, S_i=0)$ and weights $w_{0,s\bar{s}}(\mathbf{X}_i)$.

However, we do not assume that the response surface of $Y(z)$ on $\mathbf{X}_i$ is linear, and the consistency of the estimators does not rely on any modelling assumptions about $Y(z)$. Our estimators are essentially moment estimators, and their consistency and asymptotic normality follow directly from standard arguments of the law of large numbers and central limit theorem. Therefore, they have superior statistical properties compared with principal stratification analysis based on normal mixture models, which have unbounded likelihood and inaccurate asymptotic normal distribution approximations as pointed out by Frumento *et al.* (2016). Furthermore, in our simulation studies which are shown in the on-line supplementary material, we compare our method with the method of Jo and Stuart (2009) involving outcome modelling and find that our estimator is not only robust to misspecification of the outcome model but also has smaller standard error.

6. Sensitivity analysis

The theoretical foundation of the identification and estimation relies crucially on monotonicity and PI, which are fundamentally untestable. In some cases, however, these two assumptions may not be easily justified according to background knowledge. In this section, we propose approaches to conducting sensitivity analysis to assess the effect of violations of monotonicity or PI.

6.1. Sensitivity analysis for principal ignorability

The PI assumptions are critical for non-parametric identification of the PCEs as shown in Section 3. They require that the observed covariates $\mathbf{X}$ capture the key characteristics that affect both the principal stratum and the potential outcomes. They are sufficient conditions to ensure non-parametric identification, which are similar to the ignorability assumption that is used in causal inference with observational studies (Rosenbaum and Rubin, 1983a) and the sequential ignorability assumption that is used in mediation analysis (see VanderWeele (2015)). In many cases, the more covariates we observe, the more plausible these assumptions become. In practice, however, we may not be able to collect adequate covariates to remove the 'confounding' between the principal stratification and the outcome variable. Unfortunately, the PI assumptions cannot be validated by the observed data. Although there is a long history of sensitivity analysis in observational studies (e.g. Rosenbaum and Rubin (1983b) and Rosenbaum (2002)), there are only a few sensitivity analysis techniques for principal stratification analysis with binary outcomes (e.g. Sjölander *et al.* (2009) and Schwartz *et al.* (2012)) and some modelling assumptions (e.g. Gilbert *et al.* (2003)) under monotonicity. We provide a more general framework to assess the sensitivity of the deviations from the PI assumptions.

6.1.1. Under strong monotonicity

According to its proof in the on-line supplementary material, proposition 1 holds under a weaker version of PI, $E\{Y(0)|U=s\bar{s},\mathbf{X}\}=E\{Y(0)|U=\bar{s}\bar{s},\mathbf{X}\}$, which requires that the means of the control potential outcomes are the same for strata $U=s\bar{s}$ and $U=\bar{s}\bar{s}$ conditionally on covariates $\mathbf{X}$. Therefore, our sensitivity analysis is based on the deviation from this weaker assumption, captured by a single sensitivity parameter

$$\varepsilon=\frac{E\{Y(0)|U=s\bar{s},\mathbf{X}\}}{E\{Y(0)|U=\bar{s}\bar{s},\mathbf{X}\}},$$

where we implicitly assume that ε does not depend on the covariates $\mathbf{X}$. When the outcome is binary, the sensitivity parameter ε becomes the relative risk of U on the control potential

outcome $Y(0)$ given covariates $\mathbf{X}$. When $\varepsilon = 1$, the same identification results hold as those under PI. When $\varepsilon \neq 1$, we can identify the PCEs for a fixed value of ε, as shown in the following proposition.

Proposition 3. Under strong monotonicity, for a fixed value of ε, we can identify the PCEs by

$$\begin{aligned} \mathrm{ACE}_{s\bar{s}} &= E(Y|Z=1, S=1) - E\{w^{\varepsilon}_{s\bar{s}}(\mathbf{X})Y|Z=0\}, \\ \mathrm{ACE}_{\bar{s}\bar{s}} &= E(Y|Z=1, S=0) - E\{w^{\varepsilon}_{\bar{s}\bar{s}}(\mathbf{X})Y|Z=0\}, \end{aligned}$$

where

$$\begin{aligned} w^{\varepsilon}_{s\bar{s}}(\mathbf{X}) &= \frac{\varepsilon\, e_{s\bar{s}}(\mathbf{X})}{\{\varepsilon\, e_{s\bar{s}}(\mathbf{X}) + e_{\bar{s}\bar{s}}(\mathbf{X})\}\pi_{s\bar{s}}}, \\ w^{\varepsilon}_{\bar{s}\bar{s}}(\mathbf{X}) &= \frac{e_{\bar{s}\bar{s}}(\mathbf{X})}{\{\varepsilon\, e_{s\bar{s}}(\mathbf{X}) + e_{\bar{s}\bar{s}}(\mathbf{X})\}\pi_{\bar{s}\bar{s}}}. \end{aligned}$$

Although the principal scores remain the same as in proposition 1, the new weight $w^{\varepsilon}_{u}(\mathbf{X})$ further depends on the deviation from PI, with the principal score $e_{s\bar{s}}(\mathbf{X})$ overweighted by the sensitivity parameter ε. When $\varepsilon = 1$, proposition 3 reduces to proposition 1.

6.1.2. *Under monotonicity*

According to its proof in the on-line supplementary material, proposition 2 holds under a weaker version of GPI, i.e. $E\{Y(1)|U=s\bar{s}, \mathbf{X}\} = E\{Y(1)|U=ss, \mathbf{X}\}$ and $E\{Y(0)|U=s\bar{s}, \mathbf{X}\} = E\{Y(0)|U=\bar{s}\bar{s}, \mathbf{X}\}$, which require the conditional means of $Y(1)$ to be the same for strata $U=s\bar{s}$ and $U=ss$, and the conditional means of $Y(0)$ to be the same for strata $U=s\bar{s}$ and $U=\bar{s}\bar{s}$ given covariates $\mathbf{X}$. Therefore, our sensitivity analysis is based on the deviations from the above weaker assumption, captured by the following two sensitivity parameters:

$$\begin{aligned} \varepsilon_1 &= \frac{E\{Y(1)|U=s\bar{s}, \mathbf{X}\}}{E\{Y(1)|U=ss, \mathbf{X}\}}, \\ \varepsilon_0 &= \frac{E\{Y(0)|U=s\bar{s}, \mathbf{X}\}}{E\{Y(0)|U=\bar{s}\bar{s}, \mathbf{X}\}}, \end{aligned}$$

where ε_1 and ε_0 are for the potential outcomes under treatment and control respectively. The sensitivity parameters ε_1 and ε_0 enjoy transparent interpretations, which allows us to select their range according to background knowledge. For example, in the flu shot encouragement design with non-compliance that was discussed in Hirano *et al.* (2000), it may be reasonable to believe that on average the never-takers ($U=\bar{s}\bar{s}$) are the strongest patients and the always-takers ($U=ss$) are the weakest patients. In this case, the outcome of interest is an indicator of flu-related hospital visit, and therefore we can select sensitivity parameters within the range $\varepsilon_1 < 1$ and $\varepsilon_0 > 1$. We shall analyse this example in detail in Section 8.1.

For fixed values of the sensitivity parameters $(\varepsilon_1, \varepsilon_0)$, we have the following proposition.

Proposition 4. Under monotonicity, and for fixed values of $(\varepsilon_1, \varepsilon_0)$, we can identify the PCEs by

$$\begin{aligned} \mathrm{ACE}_{s\bar{s}} &= E\{w^{\varepsilon_1}_{1,s\bar{s}}(\mathbf{X})Y|Z=1, S=1\} - E\{w^{\varepsilon_0}_{0,s\bar{s}}(\mathbf{X})Y|Z=0, S=0\}, \\ \mathrm{ACE}_{\bar{s}\bar{s}} &= E(Y|Z=1, S=0) - E\{w^{\varepsilon_0}_{0,\bar{s}\bar{s}}(\mathbf{X})Y|Z=0, S=0\}, \\ \mathrm{ACE}_{ss} &= E\{w^{\varepsilon_1}_{1,ss}(\mathbf{X})Y|Z=1, S=1\} - E(Y|Z=0, S=1), \end{aligned}$$

where

$$w^{\varepsilon_1}_{1,s\bar{s}}(\mathbf{X}) = \frac{\varepsilon_1 e_{s\bar{s}}(\mathbf{X})}{\varepsilon_1 e_{s\bar{s}}(\mathbf{X}) + e_{ss}(\mathbf{X})} \Big/ \frac{\pi_{s\bar{s}}}{\pi_{s\bar{s}} + \pi_{ss}},$$
$$w^{\varepsilon_0}_{0,s\bar{s}}(\mathbf{X}) = \frac{\varepsilon_0 e_{s\bar{s}}(\mathbf{X})}{\varepsilon_0 e_{s\bar{s}}(\mathbf{X}) + e_{\bar{s}\bar{s}}(\mathbf{X})} \Big/ \frac{\pi_{s\bar{s}}}{\pi_{s\bar{s}} + \pi_{\bar{s}\bar{s}}},$$
$$w^{\varepsilon_0}_{0,\bar{s}\bar{s}}(\mathbf{X}) = \frac{e_{\bar{s}\bar{s}}(\mathbf{X})}{\varepsilon_0 e_{s\bar{s}}(\mathbf{X}) + e_{\bar{s}\bar{s}}(\mathbf{X})} \Big/ \frac{\pi_{\bar{s}\bar{s}}}{\pi_{s\bar{s}} + \pi_{\bar{s}\bar{s}}},$$
$$w^{\varepsilon_1}_{1,ss}(\mathbf{X}) = \frac{e_{ss}(\mathbf{X})}{\varepsilon_1 e_{s\bar{s}}(\mathbf{X}) + e_{ss}(\mathbf{X})} \Big/ \frac{\pi_{ss}}{\pi_{s\bar{s}} + \pi_{ss}}.$$

The weights have similar adjustments to those in proposition 3, overweighting the principal score $w_{s\bar{s}}(\mathbf{X})$ by the sensitivity parameters ε_1 and ε_0 in the treatment group and control group respectively. $\text{ACE}_{s\bar{s}}$ depends on both ε_1 and ε_0, $\text{ACE}_{\bar{s}\bar{s}}$ only on ε_0 and ACE_{ss} only on ε_1. When $\varepsilon_1 = \varepsilon_0 = 1$, proposition 4 reduces to proposition 2.

6.1.3. *Testing principal ignorability and exclusion restriction*

As a side note, proposition 4 not only allows sensitivity analysis of possible violations of GPI but also allows for testing the fundamental assumptions of GPI and the ER. To be more specific, if $\text{ACE}_{\bar{s}\bar{s}} = 0$ and $\varepsilon_0 = 1$, then

$$E(Y|Z=1, S=0) = E\{w_{0,\bar{s}\bar{s}}(\mathbf{X})Y|Z=0, S=0\}. \tag{3}$$

The contrapositive states that, if we reject condition (3) by the observed data, then we must reject $\text{ACE}_{\bar{s}\bar{s}} = 0$ or $\varepsilon_0 = 1$. Therefore, if we assume that $\text{ACE}_{\bar{s}\bar{s}} = 0$, then we can test $\varepsilon_0 = 1$; if we assume that $\varepsilon_0 = 1$, then we can test $\text{ACE}_{\bar{s}\bar{s}} = 0$. Analogous discussion applies to $\text{ACE}_{ss} = 0$ and $\varepsilon_1 = 1$. Similarly, proposition 3 implies a testable condition of PI and the ER under strong monotonicity.

Guo *et al.* (2014) proposed a parametric likelihood ratio test for GPI under monotonicity and the ER. In fact, proposition 4 implies tests for compatibility of GPI and ER, which sometimes can be an important initial step in empirical studies when we are unsure about the underlying assumptions. For instance, if we have important covariate information and impose GPI, then proposition 4 allows us to test the ER in the non-compliance setting or the causal necessity in the surrogate problem. If the test is rejected, then we may reject the ER and causal necessity, but we may also have doubts about the GPI assumption. Although this kind of discordant result does not provide a definite answer, it does warn us of the underlying assumptions and may lead us to conduct more careful analysis or better study design. See Yang *et al.* (2014) for a concrete example and philosophical discussion about checking compatibility of untestable assumptions in causal inference.

6.2. *Sensitivity analysis for monotonicity*

Without monotonicity, we have all four principal strata, and we cannot even identify their proportions without further assumptions. The inferential difficulties restrict the scope of the current literature to be under monotonicity. Some exceptions (e.g. Zhang *et al.* (2009), Ding *et al.* (2011) and Frumento *et al.* (2012)) rely on either strong modelling assumptions or additional information. We take an alternative perspective and propose an approach to performing sensitivity analysis when monotonicity is not plausible. We introduce the following sensitivity parameter ξ capturing the deviation from monotonicity:

$$\xi = \frac{\Pr(U = \bar{s}s|\mathbf{X})}{\Pr(U = s\bar{s}|\mathbf{X})},$$

which is the ratio between the probabilities of strata $U = \bar{s}s$ and $U = s\bar{s}$ conditionally on covariates $\mathbf{X}$. Furthermore, the conditional ratio is also the marginal ratio of the probabilities, i.e. $\xi = \Pr(U = \bar{s}s)/\Pr(U = s\bar{s})$. The sensitivity parameter can take values from 0 to ∞. When $\xi = 0$, we have monotonicity; when $\xi = 1$, we have equal proportions of $s\bar{s}$ and $\bar{s}s$, and thus zero average causal effect on S; when $0 < \xi < 1$, we allow deviation from monotonicity but still preserve a positive average causal effect on S; when $\xi > 1$, we have a negative average causal effect on S. Without loss of generality, we shall assume that $p_1 - p_0 \geqslant 0$ and $0 \leqslant \xi \leqslant 1$ from now on for sensitivity analysis.

Proposition 5. For a fixed sensitivity parameter ξ, we can identify the proportions by

$$\left.\begin{aligned} \pi_{s\bar{s}} &= (p_1 - p_0)/(1 - \xi), \\ \pi_{\bar{s}\bar{s}} &= 1 - p_0 - (p_1 - p_0)/(1 - \xi), \\ \pi_{ss} &= p_1 - (p_1 - p_0)/(1 - \xi), \\ \pi_{\bar{s}s} &= \xi(p_1 - p_0)/(1 - \xi), \end{aligned}\right\} \tag{4}$$

which further imply that the sensitivity parameter ξ is bounded by

$$0 \leqslant \xi \leqslant 1 - \frac{p_1 - p_0}{\min(p_1, 1 - p_0)} \leqslant 1. \tag{5}$$

Although ξ is not identifiable, the observed data provide an upper bound for it when the average causal effect on S is non-negative. Therefore, we need to perform sensitivity analysis only within the empirical version of the above bounds of ξ.

Analogously, we can show that the principal score $e_u(\mathbf{X})$ is identifiable with a known ξ, by replacing p_1 and p_0 in formula (4) by $p_1(\mathbf{X})$ and $p_0(\mathbf{X})$ respectively. Therefore, with a known ξ, GPI is sufficient to identify the PCEs.

Proposition 6. Under GPI, and for a fixed ξ, we can identify the PCEs by

$$\mathrm{ACE}_u = E\{w_{1,u}(\mathbf{X})Y|Z = 1, S = s(1)\} - E\{w_{0,u}(\mathbf{X})Y|Z = 0, S = s(0)\},$$

where $s(1)$ and $s(0)$ correspond to the values of $S(1)$ and $S(0)$ of $U = u$, and the weight $w_{z,u}(\mathbf{X})$ is defined in the same way as proposition 2.

Proposition 6 is similar to proposition 2, except that all the observed groups that are defined by (Z, S) are mixtures of two latent strata.

To end this subsection, we discuss a model strategy for principal scores without monotonicity. Combining $s\bar{s}$ and $\bar{s}s$ into one category, we define $V_i = U_i$ if $U_i = ss$ or $U_i = \bar{s}\bar{s}$, and $V_i = s\&\bar{s}$ if $U_i = s\bar{s}$ or $U_i = \bar{s}s$. We can model $\Pr(U|\mathbf{X})$ in two steps. First, we model $\Pr(V|\mathbf{X})$ as a three-level multinomial logistic regression. Second, we partition the category $s\&\bar{s}$ of V into two subcategories of $U, s\bar{s}$ and $\bar{s}s$, with probabilities $\Pr(U = s\bar{s}|V = s\&\bar{s}, \mathbf{X}) = 1/(1 + \xi)$ and $\Pr(U = \bar{s}s|V = s\&\bar{s}, \mathbf{X}) = \xi/(1 + \xi)$. We show in the on-line supplementary material the EM algorithm for computing the maximum likelihood estimate of the above model. After estimating the principal scores, we can apply the weighting and covariate adjustment method to estimate the PCEs as discussed in Section 5.

7. Simulation studies

To examine the finite sample performance of our estimators, we conduct a series of simulation studies. Let the sample sizes be 500 in all scenarios. For unit i, we generate $X_{i1}, \ldots, X_{i4} \overset{\text{IID}}{\sim} N(0, 1)$

and $X_{i5} \sim \text{Bern}(\frac{1}{2})$, and let $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{i5})^{\mathrm{T}}$. We conduct simulations under strong monotonicity and monotonicity. In each scenario we consider five cases indexed by the parameter $\theta = -1, -0.5, 0, 0.5, 1$. We postpone the interpretation of θ until afterwards.

Under strong monotonicity, for each θ we generate principal strata from a logit model $\text{logit}\{\Pr(U_i = s\bar{s}|\mathbf{X}_i)\} = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i$, where $\boldsymbol{\theta} = (0, 0.5, 0.5, 1, 1, \theta)^{\mathrm{T}}$. We generate normal potential outcomes from

$$Y_i(1)|\mathbf{X}_i \sim N\left(\sum_{j=1}^{5} X_{ij} + 2I_{\{U=s\bar{s}\}} + 1, 1\right)$$

and

$$Y_i(0)|\mathbf{X}_i \sim N\left(\sum_{j=1}^{5} X_{ij} + 2, 1\right),$$

Bernoulli potential outcomes from $\text{logit}[\Pr\{Y_i(1) = 1|\mathbf{X}_i\}] = 0.3\Sigma_{j=1}^{5} X_{ij} + I_{\{U=s\bar{s}\}}$ and $\text{logit}(\Pr[\{Y_i(0) = 1|\mathbf{X}_i\}]) = 0.3\Sigma_{j=1}^{5} X_{ij} + 0.5$.

Under monotonicity, for each θ we generate principal strata from a multinomial logit model $\Pr(U_i = u|\mathbf{X}_i) = \exp(\boldsymbol{\theta}_u^{\mathrm{T}}\mathbf{X}_i)/\Sigma_{u'}\exp(\boldsymbol{\theta}_{u'}^{\mathrm{T}}\mathbf{X}_i)$ for $u = s\bar{s}, ss, \bar{s}\bar{s}$, where $\boldsymbol{\theta}_{ss} = (0.25, 0.5, 0.5, 1, 1, \theta)$, $\boldsymbol{\theta}_{\bar{s}\bar{s}} = (-0.25, 1, 1, 0.5, 0.5, \theta)$ and $\boldsymbol{\theta}_{s\bar{s}} = \mathbf{0}$. We generate normal potential outcomes from

$$Y_i(1)|\mathbf{X}_i \sim N\left(\sum_{j=1}^{5} X_{ij} - I_{\{U=\bar{s}\bar{s}\}} + 4, 1\right)$$

and

$$Y_i(0)|\mathbf{X}_i \sim N\left(\sum_{j=1}^{5} X_{ij} + I_{\{U=ss\}} + 1, 1\right),$$

Bernoulli potential outcomes from

$$\text{logit}[\Pr\{Y_i(1) = 1|\mathbf{X}_i\}] = 0.3\sum_{j=1}^{5} X_{ij} + 0.25(I_{\{U=\bar{s}\bar{s}\}} - 1)$$

and logit outcomes

$$\Pr\{Y_i(0) = 1|\mathbf{X}_i\} = 0.3\sum_{j=1}^{5} X_{ij} + 0.25(1 - I_{\{U=ss\}}).$$

Although the above data-generating mechanisms violate GPI, they satisfy its weaker version, i.e. $\varepsilon_1 = \varepsilon_0 = 1$, which also suffices to ensure proposition 2.

To examine the performance of our estimators with and without (the weaker version of) GPI, in each simulation scenario we analyse the data with and without the binary covariate X_{i5}. Without using X_{i5}, we can view θ as a measure of the violation from GPI. In Fig. 1, we present only the results for $\text{ACE}_{\bar{s}\bar{s}}$ by using the model-assisted estimator, because in our simulations the naive weighting estimators are uniformly worse in terms of estimation efficiency. For ease of presentation, we omit similar results for other principal strata. We use 500 bootstraps to construct 95% confidence intervals and focus on the average biases and coverage rates over 1000 repeated samplings. With the binary covariate, our estimator has small biases and achieves nominal coverage rates, for both normal and Bernoulli potential outcomes. Without the binary covariate, our estimators have bias issues for both normal and Bernoulli potential outcomes when $|\theta|$ approaches 1, i.e. when GPI is severely violated. The interval estimates undercover the true parameters for normal outcomes when $|\theta|$ approaches 1, but the coverage properties for Bernoulli outcomes are robust with respect to the violations of GPI. This bias issue, as well as the untestable nature of PI and GPI, warns us that a sensitivity analysis with respect to PI

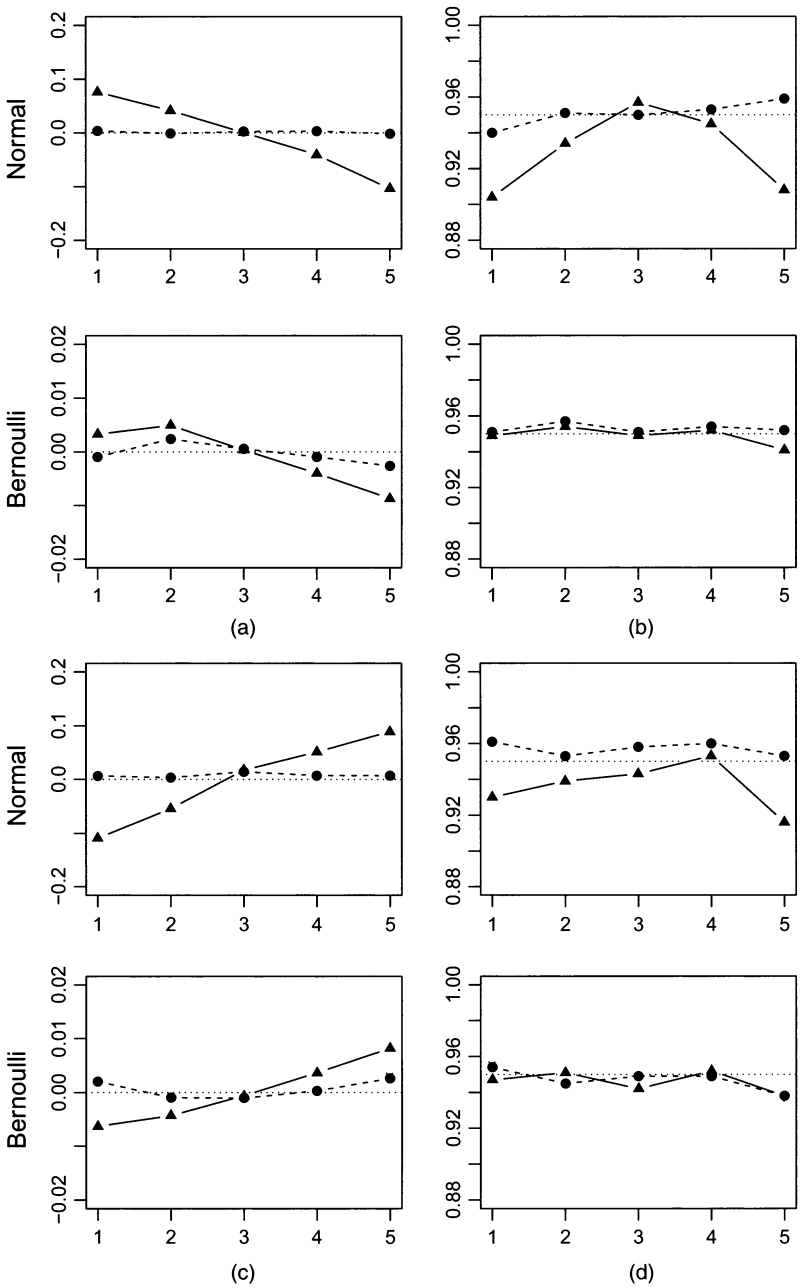

Fig. 1. Simulation results for ACE_{ss} (each subfigure is a 2 × 2 matrix summarizing two repeated sampling properties (average biases and coverage rates of interval estimates); the horizontal axis shows the case numbers; ▲, with the binary covariate; ●, without the binary covariate): (a) under strong monotonicity, bias; (b) under strong monotonicity, coverage; (c) under monotonicity, bias; (d) under monotonicity, coverage

and GPI, as proposed in Section 6.1, must be an essential part of any empirical studies using principal scores to analyse principal stratification problems.

For brevity, we compare our model-assisted estimator with the model-based estimator of Jo and Stuart (2009) in the on-line supplementary material, which shows that our estimator does not lose efficiency compared with full modelling and is robust to model misspecification of the outcome.

8. Applications

8.1. An encouragement experiment with non-compliance

In this section, we reanalyse a flu shot encouragement experiment data set that has previously been studied by Hirano *et al.* (2000). Between 1978 and 1980, a general medicine clinic in Indiana conducted an encouragement experiment, in which participating individuals' physicians were randomly assigned to the treatment arm with computer-generated letters encouraging them to inoculate their patients, or the control arm with no letters. The outcome of interest is the individual's flu-related hospitalization status during the subsequent winter. As in Hirano *et al.* (2000), we use the data from 1980, with 2893 experimental units. In our analysis, $Z=1$ if an individual's physician received the letter, and $Z=0$ otherwise. The intermediate variable $S=1$ if the individual received the flu shot, and $S=0$ otherwise. The outcome of interest $Y=1$ if the individual was hospitalized for flu-related reasons, and $Y=0$ otherwise. The monotonicity assumption is plausible for this data set, because we expect the encouragement letter to have a non-negative effect on taking the flu shot. Because this encouragement experiment is an open-label trial, previous researchers doubted the ER because of the possible 'direct effect' of the flu shot encouragement on the outcome.

To start our analysis, we use the covariate balancing conditions in corollary 2 to check the plausibility of the logistic principal score model. Choosing $h(\mathbf{X})=\mathbf{X}$ is reasonable, because all covariates are binary except for 'age'. The balance check is equivalent to estimating the PCEs on $h(\mathbf{X})$, which are known to be 0. Therefore, the corresponding 'standardized t-statistics' should follow standard normal distributions. Fig. 2(a) shows that the covariates are well balanced. Assuming GPI, we estimate the PCEs with standard errors and 95% confidence intervals in square brackets as

$$\widehat{\mathrm{ACE}}_{s\bar{s}} = -0.018\,[-0.052, 0.016],$$
$$\widehat{\mathrm{ACE}}_{ss} = -0.046\,[-0.091, 0.002],$$
$$\widehat{\mathrm{ACE}}_{\bar{s}\bar{s}} = -0.006\,[-0.030, 0.017].$$

Therefore, for compliers, receiving the encouragement letter will lower the chance of a flu-related hospital visit by 1.8%, but this effect is not significant. Furthermore, an ER seems plausible for never-takers. However, there is some evidence that it does not hold for always-takers, because the upper confidence limit is close to zero. Our findings corroborate the argument of Hirano *et al.* (2000) that 'it is more plausible to impose the exclusion restriction for never-takers than for always-takers'. Their results required careful analysis, including using data-dependent priors with several tuning parameters that account for the background knowledge. Our analysis under GPI yields coherent conclusions like theirs. Therefore, if we believe their prior knowledge and statistical analysis, then GPI seems plausible in this example. At least, there is no obvious contradiction between the two analyses, and our results under GPI have meaningful scientific interpretations.

Nevertheless, the data cannot validate GPI. It is an untestable assumption requiring that

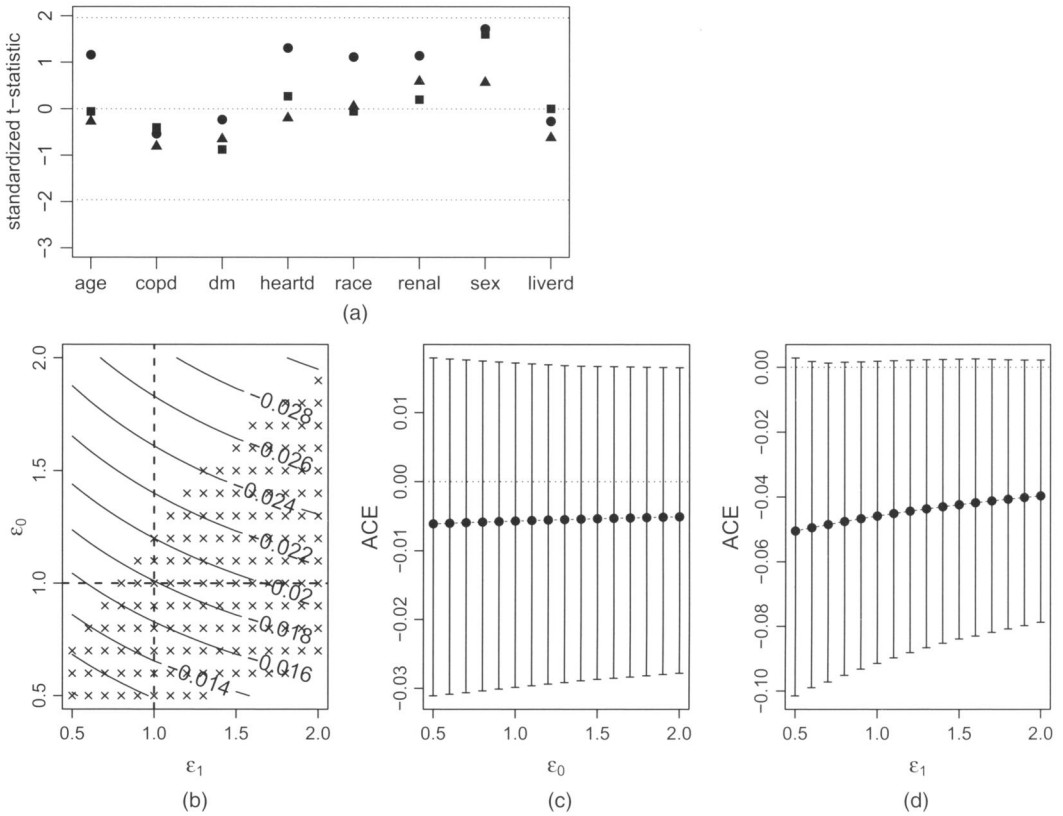

Fig. 2. Flu shot encouragement experiment: (a) covariate balance check (the horizontal axis shows the names of the covariates; ■, $U = ss$; ▲, $U = ss$; ●, $U = \bar{s}\bar{s}$); (b) sensitivity analysis for GPI, contours of the point estimates of $\text{ACE}_{s\bar{s}}$ for fixed values of ε_1 and ε_0 (×, $(\varepsilon_1, \varepsilon_0')$ such that the corresponding interval estimate covers 0); (c) sensitivity analysis for GPI, point and interval estimates of $\text{ACE}_{\bar{s}\bar{s}}$ for fixed values of ε_0; (d) sensitivity analysis for GPI, point and interval estimates of ACE_{ss} for fixed values of ε_1

observed covariates $\mathbf{X}$ contain all characteristics related to the latent principal stratum and potential outcomes. The analysis of Hirano *et al.* (2000) does not contradict GPI but does not prove it either. As advocated in Section 6, we perform sensitivity analysis for GPI, allowing $(\varepsilon_1, \varepsilon_0)$ to vary within $[\frac{1}{2}, 2] \times [\frac{1}{2}, 2]$ with results in Fig. 2(b). If we are willing to assume that never-takers are the strongest patients and the always-takers are the weakest patients, we can restrict our sensitivity analysis to within the region with $\varepsilon_1 < 1$ and $\varepsilon_0 > 1$. Interestingly, within this range most of the confidence intervals $\widehat{\text{ACE}}_{s\bar{s}}$ do not cover zero, suggesting that there is a significant causal effect for compliers. Furthermore, $\widehat{\text{ACE}}_{ss}$ and $\widehat{\text{ACE}}_{\bar{s}\bar{s}}$ are relatively robust to ε_1 and ε_0 respectively. The upper confidence limits for $\widehat{\text{ACE}}_{ss}$ are always close to zero as ε_1 varies, showing weak evidence for violation of the ER for always-takers; the centres of the confidence intervals for $\widehat{\text{ACE}}_{\bar{s}\bar{s}}$ are always close to zero as ε_0 varies, suggesting that the ER holds for never-takers. Fortunately, although the point and interval estimators vary with the sensitivity parameters, the final conclusions do not change materially.

8.2. A randomized trial with truncation by death

From October 1999 to January 2003, the Southwest Oncology Group conducted a randomized

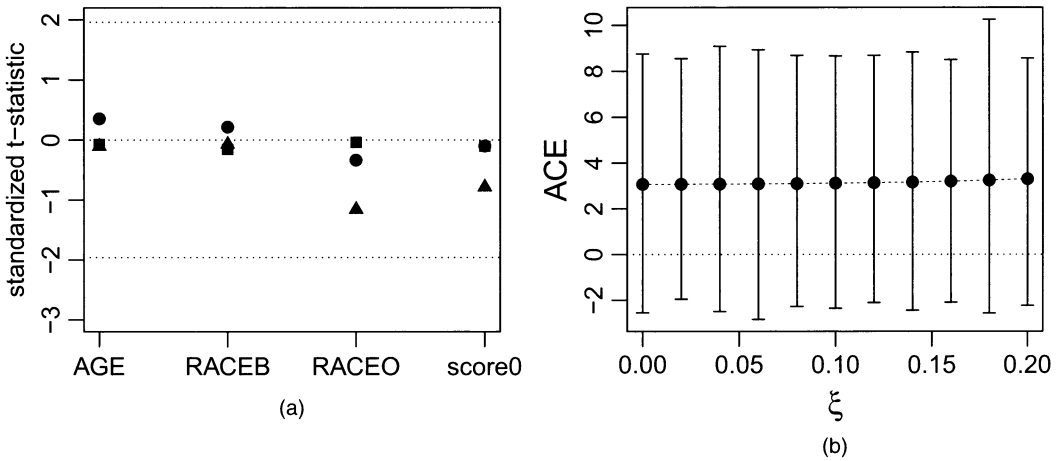

Fig. 3. Southwest Oncology Group randomized trial: (a) covariate balance check (the horizontal axis shows the names of the covariates; ■, $U = ss$; ▲, $U = ss$; ●, $U = \overline{ss}$); (b) sensitivity analysis for monotonicity (point and interval estimates of ACE_{ss} for fixed values of ξ)

phase III trial (protocol 99-16) to compare the treatment of docetaxel and estramustine (DE) with mitoxantrone and prednisone (MP) in patients with metastatic, androgen-independent prostate cancer (Petrylak *et al.*, 2004). A total of 674 eligible patients participated in the study between October 1999 and January 2003. Study participants were randomly assigned to the DE arm or the MP arm. The primary outcome is the survival time, and the secondary outcome is health-related quality of life (HRQOL). Petrylak *et al.* (2004) reported the overall survival benefit of taking DE over taking MP. In our analysis, we are interested in assessing the causal effect of DE *versus* MP on HRQOL 1 year after receiving the treatment. In our analysis, $Z = 1$ if a patient received DE, and $Z = 0$ if the patient received MP. We use the difference between HRQOL after 1 year and the baseline HRQOL as the outcome of interest. The survival indicator $S = 1$ if a patient survived after 1 year.

Because of the truncation-by-death problem, we are interested in estimating the survivor average causal effect. As in the previous example, we first check the plausibility of the logistic principal score model. Fig. 3(a) shows that we achieve covariate balance. The point estimate of the survivor average causal effect is 3.07, but its standard error is 2.976 and the 95% confidence interval $[-2.93, 8.69]$ covers zero. The results show that DE is not significantly more effective than MP to improve the HRQOL of the patients, which is similar to the results of the analysis in Ding *et al.* (2011). However, applying the normal mixture model of Zhang *et al.* (2009) we obtain point estimate 12.34 with standard error 47.17. The tremendous variability of the estimator is due to the unstable numerical issue and unreliable large sample normal distribution approximation, as investigated by Frumento *et al.* (2016).

However, both the treatment and the control are active drugs for prostate cancer, and therefore it is not reasonable to assume that the treatment is more effective than the control for all patients, i.e. monotonicity may not hold. We perform a sensitivity analysis for monotonicity and choose the range of the sensitivity parameter ξ on the basis of proposition 5. We compute from the data that $\hat{p}_1 = 0.496$ and $\hat{p}_0 = 0.389$, and therefore $0 \leqslant \hat{\xi} \leqslant 1 - (\hat{p}_1 - \hat{p}_0)/\min(\hat{p}_1, 1 - \hat{p}_0) \approx 0.217$. The sensitivity analysis results in Fig. 3(b) show that the point and interval estimates of ACE_{ss} are relatively robust to ξ. Furthermore, the interval estimates for $\widehat{\text{ACE}}_{ss}$ always cover 0 as ξ varies. In summary, the sensitivity analysis results confirm our previous conclusions.

9. Discussion

In observational studies, causal effects can be estimated by inverse propensity score weighting (Rosenbaum and Rubin, 1983a), which may be numerically unstable and have poor finite sample properties. Our estimators, weighted by probabilities themselves, do not suffer from these problems. Researchers (e.g. Bang and Robins (2005)) have developed doubly robust methods in observational studies. Similarly to our model-assisted estimators, these doubly robust estimators can also be derived from regression estimators in surveys (Cochran, 1977). Because of this similarity, it will be interesting to develop doubly robust estimators under the PI assumptions that are consistent when either the principal score or the outcome model is correctly specified.

The theoretical results have demonstrated the twofold role of the pretreatment covariates. First, the plausibility of the ignorability assumptions rely crucially on adequate covariates. Second, with more covariates that are predictive to the outcome, the covariate-adjusted estimators will be more efficient. Our results suggest that, in the design of randomized experiments, it is important for practitioners to try their best to collect covariates that are predictive of both the latent principal strata and the potential outcomes, which echos Jo and Stuart (2009) and Mealli and Pacini (2013).

Although in the main text we focused on the average causal effect within principal strata, our results can be easily extended to general causal measures. For example, we can dichotomize the outcome to identify the distributional causal effects (Ju and Geng, 2010). For binary S, we have derived explicit identification results and easy-to-implement estimators. For general discrete or continuous S, we can likewise derive theoretical results under PI by modifying the weights in propositions 2 and 5. However, a continuous S results in infinitely many principal strata, which makes it challenging to estimate the principal scores and outcome distributions conditionally on continuous variables. We need more structural assumptions on the causal problems (Jin and Rubin, 2008; Schwartz *et al.*, 2011) and more sophisticated statistical inferential tools.

Missing data are an important problem that often arises in real data analysis. Our two-step procedure has some advantages if only some outcomes are missing. We can conduct the first step for estimating principal scores without any difficulty and need to modify only the second weighting step. If the outcome is missing at random, then we can simply weight each observation by the inverse of the conditional probability of being observed given $(Z, S, \mathbf{X})$. However, for missing data problems, the key issue is the missing data mechanism. Other missing data mechanisms, e.g. latent ignorability (Frangakis and Rubin, 1999), may be more plausible, but the identification becomes challenging. Because of this complication, we leave the missing data problem for future research.

Acknowledgements

Peng Ding's work is partially supported by grant R305D150040 from Institute of Education Sciences, USA. We are grateful for the comments from Professor Donald Rubin and Professor Tirthankar Dasgupta, and other participants in the 'Matched sampling and study designs' seminar at Harvard. We benefited from the suggestions of Avi Feller at Berkeley, Keli Liu at Stanford and Professor Luke W. Miratrix and Professor Joseph K. Blitzstein at Harvard. The comments from the Joint Editor, the Associate Editor and two reviewers have helped to improve the quality of our paper significantly.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Ass.*, **91**, 444–455.

Bang, H. and Robins, J. M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–973.
Cheng, J. and Small, D. S. (2006) Bounds on causal effects in three-arm trials with non-compliance. *J. R. Statist. Soc.* B, **68**, 815–836.
Cochran, W. G. (1977) *Sampling Techniques*, 3rd edn. New York: Wiley.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.
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. *J. Am. Statist. Ass.*, **106**, 1578–1591.
Follman, D. A. (2000) On the effect of treatment among would-be treatment compliers: an analysis of the multiple risk factor intervention trial. *J. Am. Statist. Ass.*, **95**, 1101–1109.
Frangakis, C. E. and Rubin, D. B. (1999) Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, **86**, 365–379.
Frangakis, C. E. and Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics*, **58**, 21–29.
Frumento, P., Mealli, F., Pacini, B. and Rubin, D. B. (2012) Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *J. Am. Statist. Ass.*, **107**, 450–466.
Frumento, P., Mealli, F., Pacini, B. and Rubin, D. B. (2016) The fragility of standard inferential approaches in principal stratification models relative to direct likelihood approaches. *Statist. Anal. Data Minng*, **9**, 58–70.
Gilbert, P. B., Bosch, R. J. and Hudgens, M. G. (2003) Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, **59**, 531–541.
Gilbert, P. B. and Hudgens, M. G. (2008) Evaluating candidate principal surrogate endpoints. *Biometrics*, **64**, 1146–1154.
Guo, Z., Cheng, J., Lorch, S. A. and Small, D. (2014) Using an instrumental variable to test for unmeasured confounding. *Statist. Med.*, **33**, 3528–3546.
Hill, J., Waldfogel, J. and Brooks-Gunn, J. (2002) Differential effects of high-quality child care. *J. Poly Anal. Mangmnt*, **21**, 601–627.
Hirano, K., Imbens, G., Rubin, D. B. and Zhou, X.-H. (2000) Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics*, **1**, 69–88.
Hudgens, M. G. and Halloran, M. E. (2006) Causal vaccine effects on binary postinfection outcomes. *J. Am. Statist. Ass.*, **101**, 51–64.
Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: an Introduction*. New York: Cambridge University Press.
Jiang, Z., Ding, P. and Geng, Z. (2016) Principal causal effect identification and surrogate end point evaluation by multiple trials. *J. R. Statist. Soc.* B, **78**, 829–848.
Jin, H. and Rubin, D. B. (2008) Principal stratification for causal inference with extended partial compliance. *J. Am. Statist. Ass.*, **103**, 101–111.
Jo, B., MacKinnon, D. P. and Vinokur, A. D. (2011) The use of propensity scores in mediation analysis. *Multiv. Behav. Res.*, **46**, 425–452.
Jo, B. and Stuart, E. A. (2009) On the use of propensity scores in principal causal effect estimation. *Statist. Med.*, **28**, 2857–2875.
Joffe, M. M., Small, D. and Hsu, C. (2007) Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statist. Sci.*, **22**, 74–97.
Ju, C. and Geng, Z. (2010) Criteria for surrogate end points based on causal distributions. *J. R. Statist. Soc.* B, **72**, 129–142.
Mattei, A., Li, F. and Mealli, F. (2013) Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *Ann. Appl. Statist.*, **7**, 2336–2360.
Mattei, A. and Mealli, F. (2007) Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics*, **63**, 437–446.
Mattei, A. and Mealli, F. (2011) Augmented designs to assess principal strata direct effects. *J. R. Statist. Soc.* B, **73**, 729–752.
Mealli, F. and Pacini, B. (2013) Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Am. Statist. Ass.*, **108**, 1120–1131.
Petrylak, D. P., Tangen, C. M., Hussain, M. H., Lara, P. N. J., Jones, J. A., Taplin, M. E., Burch, P. A., Berry, D., Moinpour, C., Kohli, M., Benson, M. C., Small, E. J., Raghavan, D. and Crawford, E. D. (2004) Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. *New Engl. J. Med.*, **351**, 1513–1520.
Rosenbaum, P. R. (2002) *Observational Studies*, 2nd edn. New York: Springer.
Rosenbaum, P. R. and Rubin, D. B. (1983a) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rosenbaum, P. R. and Rubin, D. B. (1983b) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc.* B, **45**, 212–218.
Rubin, D. B. (1980) Comment on "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. *J. Am. Statist. Ass.*, **75**, 591–593.

Rubin, D. B. (2006) Causal inference through potential outcomes and principal stratification: application to studies with 'censoring' due to death (with discussion). *Statist. Sci.*, **21**, 299–309.
Rubin, D. B. (2007) The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statist. Med.*, **26**, 20–36.
Rubin, D. B. (2008) For objective causal inference, design trumps analysis. *Ann. Appl. Statist.*, **2**, 808–840.
Schwartz, S. L., Li, F. and Mealli, F. (2011) A Bayesian semiparametric approach to intermediate variables in causal inference. *J. Am. Statist. Ass.*, **106**, 1331–1344.
Schwartz, S., Li, F. and Reiter, J. P. (2012) Sensitivity analysis for unmeasured confounding in principal stratification settings with binary variables. *Statist. Med.*, **31**, 949–962.
Sjölander, A., Humphreys, K., Vansteelandt, S., Bellocco, R. and Palmgren, J. (2009) Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics*, **65**, 514–520.
Stuart, E. A. and Jo, B. (2015) Assessing the sensitivity of methods for estimating principal causal effects. *Statist. Meth. Med. Res.*, **24**, 657–674.
VanderWeele, T. J. (2015) *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.
Yang, F. and Small, D. S. (2016) Using post-outcome measurement information in censoring-by-death problems. *J. R. Statist. Soc.* B, **78**, 299–318.
Yang, F., Zubizarreta, J. R., Small, D. S., Lorch, S. and Rosenbaum, P. R. (2014) Dissonant conclusions when testing the validity of an instrumental variable. *Am. Statistn*, **68**, 253–263.
Zhang, J. L. and Rubin, D. B. (2003) Estimation of causal effects via principle stratification when some outcomes are truncated by 'death'. *J. Educ. Behav. Statist.*, **28**, 353–368.
Zhang, J. L., Rubin, D. B. and Mealli, F. (2009) Likelihood-based analysis of causal effects via principal stratification: new approach to evaluating job-training programs. *J. Am. Statist. Ass.*, **104**, 166–176.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Web-based supplementary material for "Principal stratification analysis using principal scores"'.