Survivor Bias and Effect Heterogeneity

Anton Strezhnev*
Draft†

April 27, 2017

Abstract

In multi-period studies where an outcome is observed at some initial point in time and at some later follow-up wave, attrition of units between periods may lead to biased estimates of causal effects in wave 2. In the presence of treatment effects on drop out, conditioning on survival may break randomization and bias estimates of causal effects. Even in the absence of bias, attrition may make estimates incomparable across waves when effects are heterogeneous across units. This paper explains how researchers can identify and estimate multi-period sub-group effects that are comparable over time in the presence of attrition. It develops a weighting strategy to adjust both for selection-induced confounding and external validity bias by equalizing the distribution of observed covariates between the always-survivor sub-group and the population at large. The paper illustrates the application of this method with a re-analysis of the Nyhan and Reifler (2015) two-wave experiment on the effect of information provision on vaccination attitudes and intent. I find that differences in sample composition are unlikely to explain the substantial changes in effects between the two waves, which suggests that results on short-term measures may be a poor indicator of the long-term effect of these information cues.

*Harvard University, Department of Government astrezhnev@fas.harvard.edu.

†Paper prepared for the 2017 New Faces in Political Methodology Conference at Penn State, April 29, 2017. The author thanks Matthew Blackwell and Marc Ratkovic for helpful comments on previous versions of this paper.

Introduction

Researchers analyzing experimental and observational datasets frequently encounter the problem of non-random sample attrition. In some situations, units may die before observation – a phenomenon common in medical studies often termed "truncation-by-death" (Zhang and Rubin, 2003). Even when death is not the truncating factor, outcomes may be defined for units conditional only on some intermediate covariate that may be affected by treatment. For example, in studies of judicial decisionmaking and the effects of panel composition on outcomes (Boyd, Epstein and Martin, 2010, e.g), a written opinion of a judge only exists for those cases where the parties fail to settle and a ruling by the judge or panel of judges is necessary.

Experimental researchers also face problems of outcome truncation. For example, many social scientists are interested in conducting survey experiments where the outcome of interest is not measured immediately after the treatment is assigned. Indeed, some experiments will wait weeks or months after treatment before eliciting a response from subjects or gather responses in two waves – an initial wave and a follow-up wave. Unless response rates are perfect, it is often the case that the resulting second-wave data suffers from non-random attrition.

Non-ignorable drop-out creates an internal validity problem as the observed data no longer identify a well-defined causal effect. If respondents in one treatment arm are less likely to remain to observation relative to respondents in another condition, analyses conditional on those units that are observed will be biased. Essentially, conditioning on the post-treatment indicator of "survival" partially undoes the original randomization of treatment. Units that were unlikely to survive may have systematically different potential outcomes than those that would always survive regardless of treatment. When treatment makes the former group more likely to survive, these units become over-represented in the treatment group relative to control.

Treatments of this problem in the statistics literature focus on strategies for estimating the causal effect for a particular "principal stratum" – a sub-population defined by its combined potential outcomes of survival under treatment and control (Frangakis and Rubin, 2002). When outcomes are truncated by attrition, Rubin (2006) argues that the only scientifically valid quantity of interest is the effect for the sub-population that would survive regardless of treatment assignment. This is because an individual causal effect is only properly defined for these particular individuals. Those who would never survive have no well-defined outcomes, while those that

would survive only under one treatment arm have no real counterfactual outcome to compare the effect of treatment. Past work has therefore focused on how to estimate this "survivor average causal effect" (SACE) when the stratum membership of units is only partially observable.

What is often ignored in the literature on principal stratification is the problem of *external validity* that can also be induced by conditioning on an intermediate variable. Even when there is zero effect of the treatment on survival and the naive estimator conditioning on survival is consistent for the true principal strata effect, the resulting estimate may still be highly misleading or uninteresting. This is because principal strata are sub-populations that may differ significantly in their underlying covariate distribution. When these covariates are sources of effect heterogeneity, principal strata effects are not comparable with effects estimated for the complete sample. When researchers have multiple waves and wish to compare effects across waves, *any* form of non-random attrition will pose a threat to reliable inference.

This paper applies to the problem of attrition bias a number of insights recently developed in the instrumental variables literature regarding the use of principal stratification methods for estimating complier treatment effects and for generalizing local causal estimands to broader populations. I outline the assumptions necessary for identification and estimation of the SACE under the assumption principal ignorability (Jo and Stuart, 2009). This assumption states that the observed covariates account for all predictors of principal stratum membership and outcome – akin to a conditional ignorability assumption in observational designs. Drawing on recent work by Ding and Lu (2016) and Feller, Mealli and Miratrix (2016) on estimation and identification of general principal stratification effects, I derive the specific principal score weighted estimator for the survivor causal effect.

I then develop an additional weighting approach that researchers can use to match the covariate distribution of the always-survivor stratum with the covariate distribution in the full sample. In the vein of Angrist and Fernandez-Val (2010) and Aronow and Carnegie (2013), this approach has the effect of down-weighting effects from units that are over-represented in the post-attrition sample and up-weighting those units that were more likely to suffer from attrition. When the observed covariates used to estimate the weights account for all of the sources of effect heterogeneity between the principal strata, effects estimated for the full sample and effects for the principal stratum can be meaningfully compared. The approach allows researchers conducting multi-wave studies with

partial attrition to account for changes in the sample composition from wave to wave and obtain comparable estimates for each time period, isolating the changes in treatment effect over time from changes in the types of units that remain under study.

I evaluate the method in a short simulation study to assess the magnitude of finite sample bias relative to conventional causal estimators for sample average effects. I then replicate an experiment by Nyhan and Reifler (2015) which examines the effect of giving individuals corrective information on the safety of the flu vaccine on individuals' beliefs that the flu vaccine is safe and their intentions to vaccinate in the future. This study was conducted in two waves, with the follow-up survey conducted about a month after the initial treatment was assigned. Unfortunately, many respondents who took the original survey could not be recontacted, resulting in substantial differences in the types of respondents who remained and those who left the study. While Nyhan and Reifler (2015) cite this attrition as a reason for not emphasizing the wave 2 results in the paper, I re-analyze their findings to show that even after correcting for changes in sample composition along a number of key covariates, many of the strong effects observed in subjects immediate responses largely dissipate in the follow-up wave.

The remainder of this paper is structured as follows: Section 2 outlines the design assumptions necessary for identification of treatment effects under attrition and presents two estimators: the principal score-weighted estimator for the survivor causal effect, and a covariate-reweighted estimator for the conditional survivor causal effect averaged over the full sample distribution of covariates. Section 3 evaluates both estimators in a small simulation study to assess the relative improvements in bias versus variance with the naive unadjusted SACE estimator. Section 4 applies the sample re-weighting method to the Nyhan and Reifler (2015) experiment and attempts to adjust for changes in covariate composition of the second wave sample. Section 5 concludes with a discussion of additional methods that could help improve the principal score estimation strategy.

Survivor Bias in Multi-period Studies

Consider a two-wave study conducted on a sample of N units. Each unit i in the sample is assigned to some treatment condition A_i. For simplicity, assume binary treatment (either 0 or 1). The wave

1 outcome Y_{i1} is observed for all units. However, not all units remain in the sample for observation in wave 2. Let S_i denote the intermediate indicator of whether a unit "survived" to be observed in the second wave.[1] Among units with $S_i = 1$, a wave 2 outcome Y_{i2} is observed. X_i denotes the observed pre-treatment covariates for unit i. For simplicity, assume discrete covariates X_i. Let $\mathcal{X}$ denote the domain of X_i. Extension of the proofs in this section to continuous covariates is straightforward, but requires a bit of additional math.

The causal quantity of interest is defined using the standard potential outcomes framework (Rubin, 1974; Holland, 1986). $Y_{i1}(a)$ and $Y_{i2}(a)$ represent the wave 1 and wave 2 outcomes that would be observed if unit i were assigned to treatment condition a. $S_i(a)$ denotes the survival status that would be observed if unit i were assigned to treatment a.

I start by considering identification of the typical quantity of interest, the Average Treatment Effect, for the first wave. Define the average treatment effect for wave t as ATE_t.

Definition 1. *Average Treatment Effect*

$$ATE_t \equiv E[Y_{it}(1) - Y_{it}(0)]$$

I make the standard assumption of consistency in order to connect the observed quantities Y_{i1} and S_i to the counterfactuals. I assume that the observed outcome for unit i having treatment $A_i = a$ equals the potential outcome if an intervention were to occur and assign that unit to treatment a. It also implies that there is a single version of treatment received by each unit and a unit's potential outcomes depend only on their own treatment assignment, often referred to as the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1990).

Assumption 1. *Consistency*

$$\begin{aligned} Y_{i1} &= A_i \cdot Y_{i1}(1) + (1 - A_i) \cdot Y_{i1}(0) \\ S_i &= A_i \cdot S_i(1) + (1 - A_i) \cdot S_i(0) \\ Y_{i2} &= A_i \cdot Y_{i2}(1) + (1 - A_i) \cdot Y_{i2}(0) \end{aligned}$$

[1] Throughout this paper, I will use "survival" as shorthand to denote whether units are able to be observed in the follow-up wave – while survival is literally the problem in the case of "truncation-by-death," non-survival can also refer to situations where the respondent exists, but is unable to have an outcome response elicited by the researcher.

Second, I assume that treatment assignment is independent of all three sets of potential outcomes

Assumption 2. *Ignorability*

$$Y_{i2}(a), Y_i1(a), S_i(a) \perp\!\!\!\perp A_i| \ \forall a \in \{0,1\}$$

In experimental settings the random assignment of treatment is sufficient to guarantee that ignorability holds unconditionally. For observational designs, it is necessary for researchers to make a weaker assumption and condition on any observed pre-treatment confounding factors X_i – the conventional "selection on observables" assumption.

Assumption 3. *Conditional Ignorability*

$$Y_{i2}(a), Y_i1(a), S_i(a) \perp\!\!\!\perp A_i|X_i \ \forall a \in \{0,1\}$$

For generality, I will discuss identification and estimation assuming conditional ignorability, but in cases where stronger ignorability assumptions hold, one can omit certain elements such as propensity of treatment weighting from the estimators.

Define the stabilized inverse propensity of receiving treatment a for unit i (Robins, 1986; Robins, Hernan and Brumback, 2000).

Definition 2. *Stabilized Inverse Propensity Score*

$$\frac{1}{\pi_i^a} = \frac{Pr(A_i = a)}{Pr(A_i = a|X_i = x)}$$

Finally, I make a "positivity" or "overlap" assumption, that the probability of receiving treatment is between 0 and 1 for all covariate profiles X_i (Hernan and Robins, 2017).

Assumption 4. *Positivity*

$$0 < Pr(A_i = a) < 1|X_i \ \forall i = \{1, \ldots, N\}$$

Under this standard set of assumptions 1-3, the ATE for wave 1 outcomes is identifiable and

the standard inverse probability of treatment weighted difference-in-means estimator between the treated and control groups is consistent for it (Imbens, 2004).

However, for second wave outcomes, the average treatment effect is undefined as some units in the sample do not have well-defined individual causal effects. Truncation of outcomes by attrition limits the set of potential outcomes that are well-defined. Following (Zhang and Rubin, 2003), I consider that $Y_{i2}(a)$ is defined as only for $S_i(a) = 1$. The difference between $Y_{i2}(1)$ and $Y_{i2}(0)$ is real-valued only for units that would survive under both treatment conditions. For those that would never respond, no coherent causal quantity can be defined (without additional interventions). Likewise, for units that survive under the factual treatment, but would not survive under the counterfactual, one of the two potential outcomes is undefined.

Using the principal stratification approach of Frangakis and Rubin (2002), I segment the sample of units into four "principal strata" based on the joint potential survival outcomes $\{S_i(1), S_i(0)\}$. Table 1 describes the principal strata that exist under binary treatment and intermediate variables (using the terminology of (Frangakis and Rubin, 2002)) and the potential outcomes that exist and are observed under each treatment arm. Note that these strata are *pre-treatment* quantities as treatment by definition cannot affect to which stratum a unit belongs, only the value of S_i that is observed. Therefore, causal quantities can be properly defined as averages conditional on a specific stratum membership.[2] Zhang and Rubin (2003) argue that when potential outcomes are truncated (undefined) when the intermediate variable is S_i is 0, the only group for which a well-defined individual treatment effect exists is the sub-population that would always be observed at follow-up in wave 2 under either treatment condition, the "always-survivor" stratum with $S_i(1) = S_i(0) = 1$.

Note that the distinction between "missingness" and "truncation" is important. A "missing" value is one that is realized but unobserved by the researcher while a "truncated" value is one that is unrealized (Zhang and Rubin, 2003). While some researchers may consider drop-out in surveys to be an example of missingness (as respondents' underlying attitudes still exist in some sense), it is arguably more appropriate to treat it as truncation. The actual event of the respondent answering the question never comes to pass for respondents who drop-out. In order to observe the

[2]Note that other stratum effects can be of interest depending on the design. Instrumental variables approaches estimate effects for the principal stratum in which units' intermediate S_i matches the assigned value of A_i (the "complier" group)(Angrist, Imbens and Rubin, 1996)

$S_i(1)$	$S_i(0)$	Stratum name	A_i	$Y_{i1}(1)$	$Y_{i1}(0)$	$Y_{i2}(1)$	$Y_{i2}(0)$
1	1	Always Survivors (LL)	1	Y_{i1}	--	Y_{i2}	--
			0	--	Y_{i1}	--	Y_{i2}
1	0	Protected (LD)	1	Y_{i1}	--	Y_{i2}	--
			0	--	Y_{i1}	$*$	$*$
0	1	Harmed (DL)	1	Y_{i1}	--	$*$	$*$
			0	--	Y_{i1}	--	Y_{i2}
0	0	Never Survivors (DD)	1	Y_{i1}	--	$*$	$*$
			0	--	Y_{i1}	$*$	$*$

Table 1: Definition of Principal Strata from Frangakis and Rubin (2002)

value of the outcome, it is not enough to change the missingness mechanism such that the latent outcome is "uncovered." Observing the outcome would require considering some post-treatment intervention on the non-respondents that would force them to take the survey. This ultimately changes the quantity of interest by altering the intervention being evaluated from the simple treatment to the treatment combined with a "survival intervention." Since there are arguably many such post-treatment interventions each with different effects on the outcome, the causal effect of interest may, as a result, be poorly defined (Cole and Frangakis, 2009). The application of principal stratification therefore goes beyond the limited set of cases where units cease to exist post-treatment and is a useful theoretical concept for almost any situation where missingness results from post-treatment events that block units from realizing the outcome.

For the second wave, treatment effects are only identifiable for the LL or "always-survivor" stratum. Denote the average effect for wave t this sub-population as the Survivor Average Causal Effect (SACE) (Zhang and Rubin, 2003).

Definition 3. *Survivor Average Causal Effect (SACE)*

$$SACE_t \equiv E[Y_{it}(1) - Y_{it}(0) | S_i(1) = S_i(0) = 1]$$

Conditioning on the principal stratum does not induce post-treatment bias. However, stratum membership is not perfectly observed for each unit. What is observed is the post-treatment quantity S_i which constrains a unit to being a member of one of two strata. Table 2 illustrates the connection between observed quantities and principal strata membership.

While all units with $S_i = 0$ can be ruled out of the always-survivor stratum, the set of units with

S_i	A_i	Possible Strata
1	1	Always Survivor (LL) or Protected (LD)
1	0	Always Survivor (LL) or Harmed (DL)
0	1	Never Survivor (DD) or Harmed (DL)
0	0	Never Survivor (DD) or Protected (LD)

Table 2: Definition of Principal Strata from Frangakis and Rubin (2002)

$S_i = 1$ is a mixture of always-survivors and protected/harmed units. Taking average differences in outcome between treated and control arms conditional on observed survival status may be biased for the SACE since controlling for the intermediate S_i induces a form of post-treatment selection bias, *even* in the case where treatment is fully randomly assigned.

Proposition 1. *Under assumptions 1 2, and 4, the bias of the naive difference-in-means estimator* $\widehat{SACE}_t^{IPW}$ *for the true* $SACE_t$ *is*

$$E[\widehat{SACE}_t^{IPW}] - SACE_t = (E[Y_{it}(1)|S_i(1) = 1, S_i(0) = 0] - E[Y_{it}(1)|S_i(1) = 1, S_i(0) = 1]) \frac{\pi_{10}}{\pi_{11} + \pi_{10}} - (E[Y_{it}(0)|S_i(1) = 0, S_i(0) = 1] - E[Y_{it}(1)|S_i(1) = 1, S_i(0) = 1]) \frac{\pi_{01}}{\pi_{11} + \pi_{01}}$$

where $\pi_{ab} = Pr(S_i(1) = a, S_i(0) = b)$

As intuition would suggest, the bias for the SACE is 0 if there is no individual level effect of treatment on survival, that is $S_i(1) = S_i(0) \forall i$. Under the assumption of zero individual causal effect, there exist only two principal strata, the "always survivors" and the "never survivors," which are perfectly identified by the observed survival status.

Alternately, even in the presence of some non-random attrition, if the expected potential outcomes in the always-survivor and partial survivor groups are equal, then the naive difference-in-means estimator remains unbiased for the true effect. Unfortunately for researchers, randomization of treatment does not guarantee "quasi-randomization" of principal strata. Assuming that the potential outcome distributions do not differ across principal strata is a very heroic assumption without additional conditioning covariates. For example, in medical studies where subjects may die prior to observation, it is often the case that individuals who would be saved by treatment are likely to be generally less healthy on average than those who would survive under both treatment and control. As a result, the distribution of latent health among treated and control units will no

longer be balanced.

Point identification of the SACE is not possible without additional assumptions as the strata are only partially observed. Sharp non-parametric bounds on the treatment effect can be obtained and refined using additional assumptions on the distribution of principal strata (Zhang and Rubin, 2003; Imai, 2008). These assumptions can be further augmented by assuming a parametric model for the stratum proportions and using a fully Bayesian approach to estimation (Zhang, Rubin and Mealli, 2009). Other approaches have suggested identification using auxiliary covariates similar to instrumental variables (Ding et al., 2011; Mattei, Mealli and Pacini, 2014). Recently, Tchetgen Tchetgen et al. (2012) and Tchetgen Tchetgen (2014) proposed a weighting method incorporating post-treatment covariates for point identification.

This paper presents identification and estimation results drawing on recent work on "principal score" methods for estimating principal strata effects – a close analogue of propensity score methods. This approach was proposed by Jo and Stuart (2009) for generally estimating effects for sub-groups defined by the potential outcomes of some post-treatment variable. The paper primarily dealt with causal effects within strata defined by treatment compliance. In that context, while effects are technically defined for all strata, the researcher's stratum of interest is typically the group that would be moved by treatment assignment to comply with the treatment. Borrowing the terminology of Angrist, Imbens and Rubin (1996), the goal is to estimate the effect for the "complier" stratum rather than the "always-takers" or "never-takers." While the stratum of interest in the treatment compliance case differs from that of the survivor bias case, the general concept of principal score weighting can nevertheless be applied. Ding and Lu (2016) and Feller, Mealli and Miratrix (2016) developed the general theoretical framework required for identification of principal strata effects using covariate adjustment and highlighted a potential application to truncation-by-death. This paper draws on those earlier results to outline the necessary assumptions for estimating principal scores for the "always-survivor" stratum. It then extends the estimation strategy to allow researchers to adjust principal strata effect estimates to different covariate distributions, utilizing analogous results from the literature on instrumental variables and extrapolation from Local Average Treatment Effects (LATEs) (Angrist and Fernandez-Val, 2010).

The "principal score" is defined as the probability of a unit appearing in a principal stratum conditional on some set of covariates. These covariates serve the function of "confounding"

variables in a conventional selection-on-observables identification strategy, except instead of predicting the treatment and outcome, they confound principal stratum membership and outcome. Re-weighting by the principal score has the effect of matching the distribution of observed covariates among the observed survivors with the distribution of covariates in the always-survivor stratum. In this sense, "principal score" methods rely on a similar intuition as propensity score methods for confounding adjustment (Rosenbaum and Rubin, 1983).

Identification of the principal strata proportions and the principal score for the always survivors requires an assumption of monotonicity. This is common to many previous approaches to identification of the SACE (Zhang and Rubin, 2003; Zhang, Rubin and Mealli, 2009, e.g.). It states that for all units in the sample, the individual effect of treatment on survival is strictly positive or strictly negative. Without loss of generality, I assume that no units would drop out under treatment if they would not drop out under control (if the converse is the case, then one could simply relabel treated and control to satisfy the assumption). As in the instrumental variables context (Angrist, Imbens and Rubin, 1996), this assumption allows researchers to rule out one of the principal strata and identify the remaining stratum memberships using only observed survival proportions under treatment and control.

Assumption 5. *Monotonicity*

$$Pr(S_i(1) \geq S_i(0)) = 1$$

Under monotonicity, all units with $A_i = 0$, $S_i = 1$ are in the always-survivor stratum while the treated survivors are a combination of always survivors and protected units.

Proposition 2. *Principal stratum proportions*

Under monotonicity, the stratum proportions are

$$P(S_i(1) = S_i(0) = 1) = Pr(S_i(0) = 1)$$

$$P(S_i(1) = 1, S_i(0) = 0) = Pr(S_i(1) = 1) - Pr(S_i(0) = 1)$$

$$P(S_i(1) = 1, S_i(0) = 0) = 0$$

and

$$P(S_i(1) = 0, S_i(0) = 0) = 1 - Pr(S_i(1) = 1)$$

which can be estimated by

$$\widehat{Pr(S_i(a) = 1)} = \frac{1}{\sum_{i:A_i=a} \frac{1}{\hat{\pi}_i^a}} \sum_{i:A_i=a} \frac{1}{\hat{\pi}_i^a} S_i$$

where $i : A_i = a$ *denotes the sum over all units with treatment* a *and* $\frac{1}{\hat{\pi}_i^a}$ *is the estimated inverse propensity of treatment weight for observation* i*.*

Identifying the stratum proportions also allows the identification of covariate means and distributions in the always survivor stratum using an approach similar to kappa weighting for compliers in the instrumental variables set-up Abadie (2003). This allows researchers to assess the degree to which the stratum of interest differs from the population at large.

Proposition 3. *Conditional covariate expectations in the always-survivor stratum*

$$E[X_i|S_i = S_i = 1] = \frac{E[\gamma(x)X_i]}{E[\gamma(x)]}$$

where

$$\gamma(x) = \frac{(S_i(1 - A_i) \times (1 - P(A_i = 1))}{1 - P(A_i = 1|X_i)}$$

Intuitively, all observations in the $A_i = 0, S_i = 1$ group belong to the always-survivor principal stratum. If treatment is made ignorable from X_i (by reweighting the sample by inverse propensity of treatment), then $E[X_i|A_i = 0, S_i(1) = S_i(0) = 1] = E[X_i|A_i = 1, S_i(1) = S_i(0) = 1] = E[X_i|S_i(1) = S_i(0) = 1]$

I then define the (conditional) principal score for unit i as the probability that a unit is in the always-survivor stratum conditional on covariates and observed treatment and survival status.

Proposition 4. *Principal Score*

Under assumptions 1, 3, 4, and 5

$$\widehat{w_i^a} = \frac{Pr(S_i = 1|\widehat{A_i = 0}, X_i = x)}{Pr(S_i = 1|\widehat{A_i = a}, X_i = x)}$$

is a consistent estimator of

$$Pr(S_i(1) = 1, S_i(0) = 1|X_i = x, A_i = a, S_i = 1)$$

The principal score is 1 for units under control that survive and a ratio of conditional survival probabilities between treated and control groups. Identification of the SACE requires one final assumption of *weak principal ignorability* (Feller, Mealli and Miratrix, 2016).

Assumption 6. *Weak Principal Ignorability*

$$Y_i(1) \perp\!\!\!\perp S_i(0)|X_i, S_i(1)$$

Under observed S_i, this amounts to making a cross-world independence assumption that the always survivors and protected have the same potential outcomes conditional on treatment and background characteristics. A stronger version of this assumption assumes that all stratum memberships are ignorable conditional on X_i. In this case, it is sufficient to assume ignorability between only the always-survivor and protected strata.

Unfortunately, these ignorability assumptions cannot be guaranteed even in a randomized experiment since stratum membership is not manipulable by the researchers. Post-treatment selection effectively changes an experimental setting into an observational design. Researchers should therefore be cautious when estimating principal strata effects and attempt to measure a rich set of possible stratum confounders. Sensitivity analyses can also be used to assess the magnitude of unobserved confounding that would substantially alter an estimate that assumes ignorability (Blackwell, 2014).

With these two additional assumptions, the SACE can be identified and estimated using an inverse-propensity weighting approach

Proposition 5. *Under assumptions 1, 4, 3, 5, and 6* $E[\widehat{SACE_t^{PS}}]$ *is consistent for the SACE.*

$$E[\widehat{SACE_t^{PS}}] = \frac{1}{\sum_{i:A_i=1,S_i=1} \frac{\hat{w_i^1}}{\hat{\pi_i^1}}} \sum_{i:A_i=1,S_i=1} \frac{\hat{w_i^1} Y_{it}}{\hat{\pi_i^1}} - \frac{1}{\sum_{i:A_i=0,S_i=1} \frac{1}{\hat{\pi_i^0}}} \sum_{i:A_i=a,S_i=1} \frac{Y_{it}}{\hat{\pi_i^0}}$$

where $\sum_{i:A_i=a,S_i=1}$ *denotes the sum over all units i with treatment a and survival status* 1*.*

Principal score re-weighting adjusts the covariate distribution of treated survivors to match the distribution of survivors assigned control, which are always-survivors under monotonicity. If the covariates explain all heterogeneity in potential outcomes between always-survivors and the protected strata, matching the covariate distributions suffices to identify the SACE.

In the absence of individual treatment effects on survival, the SACE is identified without bias even when conditioning on the post-treatment indicator for survival. However, even though the estimator of the SACE is consistent, conditioning on survival changes the distribution over which the quantiy of interest is defined. The SACE is a *sub-group* effect, meaning that heterogeneity in treatment effects can induce external validity problems if the always-survivor sub-group is particularly unrepresentative of the population of interest. This can create interpretability issues in multi-wave studies when comparing estimated effects over time. While causal effects for wave 2 will be *internally valid* they will not be *externally valid* in the sense that they are not defined for the same sub-population as effects estimated for the first wave. While a researcher could change the quantity of interest to compare the SACE for both wave 1 and wave 2, this would be effectively throwing away data in estimating the wave 1 effect by eliminating from that wave all observations that suffered attrition.

Moreover, for many types of interventions, researchers want to know the effect for a set of units that is well-defined by factors that exist outside of the context of the individual study (e.g. the overall population of a country). This can be important for replicability of research as fixing the population of interest is necessary to compare estimates across studies. However, the SACE is defined with respect to a sub-group that is unique to the particular study being conducted. If the hypothetical study were conducted in a manner that changed attrition rates, such as by investing more resources into re-contacting respondents, the SACE may be different for identical treatments.

One approach to addressing the external validity problem is to re-weight the always-survivors

such that their covariate distributions match some target distribution, such as the sample distribution. This strategy draws on insights from the literature on Local Average Treatment Effect (LATE) estimation. It is well known that instrumental variables designs identify treatment effects only for a subset of observations "encouraged" to adopt treatment by the instrument (Angrist, Imbens and Rubin, 1996). In many cases, this local treatment effect can differ from the true average treatment effect. To overcome this problem, inverse propensity weighting estimators have been suggested as a method for obtaining more generalizable estimates of treatment effects (Angrist and Fernandez-Val, 2010; Aronow and Carnegie, 2013). Intuitively, observations that are comparatively more likely to appear in the principal stratum relative to the sample at large receive lower weights while those that are under-represented receive higher weights.

Identification of principal scores and principal strata distributions allows the application of a similar strategy for re-weighting the SACE. Define the re-weighted SACE (RWSACE) as the average of SACEs within covariate strata averaged over the distribution of those strata in some target population (such as the sample).

Definition 4. *Re-weighted Survivor Average Causal Effect*

$$PWSACE_t \equiv \sum_{x \in \mathcal{X}} E[Y_{it}(1) - Y_{it}(0)|X_i = x]Pr(X_i = x)$$

Define the principal stratum sample selection weight as

Definition 5. *Inverse sample selection weight Under assumptions 1*

$$q_i = \frac{Pr(S_i = 1|A_i = 0)}{Pr(S_i = 1|A_i = 0, X_i = x)}$$

Proposition 6. *Under assumptions 1, 3, 4, 5, and 6*

The principal score and selection-weighted estimator

$$\widehat{RWSACE}_t^{PS} = \frac{1}{\sum_{i:A_i=1,S_i=1} \frac{\hat{q}_i \hat{w}_i^1}{\hat{\pi}_i^1}} \sum_{i:A_i=1,S_i=1} \frac{\hat{q}_i \hat{w}_i^1 Y_i}{\hat{\pi}_i^1} - \frac{1}{\sum_{i:A_i=0,S_i=1} \frac{\hat{q}_i}{\hat{\pi}_i^0}} \sum_{i:A_i=a,S_i=1} \frac{\hat{q}_i Y_i}{\hat{\pi}_i^0}$$

is consistent for the re-weighted Survivor Average Causal Effect.

Simulation Study of Finite Sample Performance

While the estimators above are consistent for the true causal effects, the estimated weights are ratio estimators and therefore biased in small samples. To assess the bias in the proposed principal score estimators relative to standard ATE estimators in finite samples and establish some guidelines for the sample sizes necessary for reliable estimation, I conduct a small simulation study.

For simplicity, I assume a single binary covariate X_i, that is both a confounder of treatment and principal strata. X_i is also assumed to be the only source of treatment effect heterogeneity. For ease of interpretability, I generate an outcome variable Y_i for all units (so the ATE is defined), and focus on differences in effects between the overall sample and the survivor principal stratum.[3].

The specific data-generating process I used is:

$$
\begin{aligned}
X_i &\sim \text{Bernoulli}(.5) \\
A_i|X_i &\sim \text{Bernoulli}(.3 \times \mathcal{I}(X_i = 0) + .6 \times \mathcal{I}(X_i = 1)) \\
Pr(S_i(1) = S_i(0) = 1|X_i) &= .6 \times \mathcal{I}(X_i = 0) + .4 \times \mathcal{I}(X_i = 1) \\
Pr(S_i(1) = 1, S_i(0) = 0|X_i) &= .2 \times \mathcal{I}(X_i = 0) + .5 \times \mathcal{I}(X_i = 1) \\
Pr(S_i(1) = S_i(0) = 0|X_i) &= .2 \times \mathcal{I}(X_i = 0) + .1 \times \mathcal{I}(X_i = 1) \\
Y_i|A_i, X_i &= 1 + 2A_i + 8X_i - 16A_iX_i + \varepsilon_i \\
\varepsilon_i &\sim \mathcal{N}(0, 12)
\end{aligned}
$$

where $\mathcal{I}(X_i = a)$ is an indicator variable that takes on 1 if $X_i = a$ and 0 otherwise and $\mathcal{N}(\mu, \sigma^2)$ denotes a draw from a normal distribution with mean μ and variance σ^2.

Table 3 displays the bias, variance and mean-squared error of the two estimators of the ATE – the standard IPW difference-in-means and the re-weighted SACE. Table 4 reports the bias, variance and MSE for the naive (conditional on survival) and principal score weighting estimators of the SACE. As covariate and treatment are discrete, all models in this simulation were estimated non-parametrically using the relevant conditional means. I run 5000 iterations for each of the six selected sample sizes from 75 to 2400.

[3]This can be interpreted as a two-wave study with zero difference in average treatment effects between waves, but with attrition in the second wave

	ATE: Difference-in-means			ATE: Re-weighted SACE		
Sample Size	Bias	Variance	MSE	Bias	Variance	MSE
75	-0.004	0.737	0.737	0.001	1.172	1.171
150	0.002	0.371	0.371	0.000	0.578	0.578
300	-0.001	0.185	0.185	-0.003	0.280	0.280
600	-0.005	0.088	0.088	-0.005	0.137	0.137
1200	-0.003	0.046	0.046	-0.002	0.069	0.069
2400	-0.001	0.023	0.023	0.000	0.034	0.034

Table 3: Simulation study of ATE estimators

	SACE: Naive (conditional on survival)			SACE: Principal score weighting		
Sample Size	Bias	Variance	MSE	Bias	Variance	MSE
75	1.259	1.139	2.724	-0.042	1.932	1.934
150	1.255	0.558	2.132	-0.019	0.927	0.927
300	1.250	0.269	1.832	-0.014	0.447	0.447
600	1.246	0.137	1.690	-0.010	0.226	0.226
1200	1.246	0.067	1.620	-0.004	0.110	0.110
2400	1.251	0.034	1.600	-0.002	0.054	0.054

Table 4: Simulation study of SACE estimators

In the presence of confounding, the naive SACE estimator is biased and this bias does not diminish as the sample size gets larger. The principal score-weighted SACE estimator sacrifices increased variance for a significant reduction in bias. Notably though, there exists some finite sample bias in the principal score-weighted estimator. However, this bias becomes negligible for moderate sample sizes (around 200 to 300). Certainly, the rate at which this bias decays will vary depending on the magnitude of confounding and the complexity of the true selection model, but the initial analysis suggests that a few hundred observations is sufficient to achieve asymptotic unbiasedness.

Interestingly, the reweighted SACE exhibits much less small-sample bias and is in general less variable than the principal-score weighted SACE. This is in part because multiplying the sample selection weight by the conditional principal score cancels out the estimator $Pr(S_i = 1|\hat{A_i} = 0, S_i = 1)$ making the weights no longer ratio estimates. This also seems to generate lower variance in the estimator. If the specific always-survivor stratum is not of particular theoretical interest for researchers, then the average of always-survivor effects across different covariate distributions may be a reasonable alternative quantity of interest that can be estimated with greater precision.

Application: Nyhan and Reifler (2015) Vaccine Messaging Experiment

This section further illustrates the utility of the weighting estimators developed in this paper by applying them to an existing empirical study. Nyhan and Reifler (2015) report a survey experiment designed to test the effect of corrective information on individuals' beliefs in the safety of the flu vaccine and their intention to get vaccinated.

The study was conducted in two waves of the 2012 Cooperative Congressional Election Survey (CCES) in October and November of 2012. 1000 initial respondents were randomly assigned to be exposed to one of three messages in the first wave: a "danger" message describing the risks of influenza, a "corrective" message explaining that the flu vaccine does not cause the flu, or a "control" group with no message. Respondents were then asked three questions: whether they believe the flu vaccine causes the flu, whether they believe the flu vaccine is safe, and whether they intend to get vaccinated in the coming year. Respondents were exposed to the message only in the first wave of the survey. Table 5 replicates the paper's reported effect estimates of each of the two message treatments relative to the control on outcomes in wave 1 for the entire sample and in two sub-groups defined by individuals' pre-treatment levels of concern over vaccine side-effects (high and low).[4]

In the second wave, re-contacted respondents were asked the same three outcome questions but not re-exposed to the treatment. The intent of conducting the survey in two waves was to assess both the immediate effects of the treatment and potential longer-term effects. Table 6 reports the naive estimates of the SACE using the simple difference-in-means among wave 2 respondents only. Most of the effects of the correction on misperceptions over vaccine safety appear to have dissipated in the follow-up period. High concern individuals assigned to the correction are also not significantly more likely to report lower vaccination intent in wave 2. Interestingly, within the sub-group defined by high pre-treatment concern about vaccine side-effects, there appears to be

[4]Because the outcomes are measured as ordered Likert-scale responses, Nyhan and Reifler (2015) report ordered probit coefficient estimates. However, these coefficients are not directly interpretable in terms of the outcome variable. For simplicity, I report results as estimates from an OLS model (differences-in-means). The original paper also conducted analyses using OLS which were substantively identical to those from the ordered probit. Applying both OLS and latent variable models to ordinal outcomes requires additional modeling assumptions to recover valid causal effects (Volfovsky, Airoldi and Rubin, 2015). Indeed, the ATE itself is ill-defined for non-numeric outcomes

Subgroup	Treatment	Vaccine gives flu (1-4)	95% CI	N
Full sample	Danger	0.062	$[-0.157, 0.287]$	655
	Correction	-0.371^{*}	$[-0.613, -0.122]$	660
Low concern	Danger	0.063	$[-0.176, 0.321]$	504
	Correction	-0.301^{*}	$[-0.553, -0.036]$	513
High concern	Danger	-0.006	$[-0.392, 0.383]$	151
	Correction	-0.478^{*}	$[-0.906, -0.014]$	147

Subgroup	Treatment	Flu vaccine unsafe (1-4)	95% CI	N
Full sample	Danger	-0.114	$[-0.303, 0.090]$	658
	Correction	-0.227^{*}	$[-0.406, -0.036]$	660
Low concern	Danger	-0.076	$[-0.270, 0.118]$	507
	Correction	-0.196^{*}	$[-0.390, -0.017]$	514
High concern	Danger	-0.287	$[-0.679, 0.127]$	151
	Correction	-0.124	$[-0.504, 0.260]$	146

Subgroup	Treatment	Intend to vaccinate (1-6)	95% CI	N
Full sample	Danger	0.224	$[-0.262, 0.686]$	659
	Correction	0.002	$[-0.458, 0.476]$	660
Low concern	Danger	0.332	$[-0.239, 0.827]$	508
	Correction	0.117	$[-0.448, 0.738]$	513
High concern	Danger	-0.038	$[-1.000, 0.874]$	151
	Correction	$-0.837^{\dagger}$	$[-1.716, 0.022]$	147

Observations weighted by CCES sample weights. Sample sizes are the sum of units in the relevant treatment condition and in the no-message control.
95% percentile confidence intervals obtained from 1000 bootstrap resamples.
$* = p < .05, \dagger = p < .1$.

Table 5: Nyhan and Reifler (2015) – Estimated average treatment effects – Wave 1

a "backfire" effect where high concern respondents assigned to read the correction in October are more likely to think that the flu vaccine is unsafe when asked in November. Note however that the estimate is only significant at the $p < .1$ level and it is important to be cautious about interpreting high-variance effects particularly for small sub-groups. Nevertheless, the existence of a backfire effect is consistent with other experiments on the effect of corrections (Nyhan and Reifler, 2010, e.g).

However, the authors chose to not report analyses for wave 2 in the main section of the paper as they found that their sample suffered from significant attrition between waves. Of the 1000 respondents interviewed in the first wave, 178 could not be re-contacted. They note that the distributions of observed pre-treatment covariates differ substantially between the sample at large and the subset that could be recontacted.

While the authors cite attrition bias as the reason for downplaying the second wave effects, it

Subgroup	Treatment	Vaccine gives flu (1-4)	95% CI	N
Full sample	Danger	0.096	$[-0.137, 0.371]$	535
	Correction	-0.195	$[-0.442, 0.075]$	542
Low concern	Danger	0.059	$[-0.212, 0.324]$	421
	Correction	-0.107	$[-0.423, 0.173]$	434
High concern	Danger	0.010	$[-0.435, 0.501]$	114
	Correction	-0.418	$[-0.953, 0.120]$	108

Subgroup	Treatment	Flu vaccine unsafe (1-4)	95% CI	N
Full sample	Danger	-0.016	$[-0.239, 0.205]$	536
	Correction	0.096	$[-0.117, 0.298]$	543
Low concern	Danger	-0.020	$[-0.231, 0.198]$	421
	Correction	-0.030	$[-0.253, 0.186]$	434
High concern	Danger	0.249	$[-0.241, 0.684]$	115
	Correction	$0.416^{\dagger}$	$[-0.046, 0.837]$	109

Subgroup	Treatment	Intend to vaccinate (1-6)	95% CI	N
Full sample	Danger	0.259	$[-0.372, 0.798]$	536
	Correction	0.025	$[-0.479, 0.594]$	543
Low concern	Danger	0.303	$[-0.309, 0.958]$	421
	Correction	0.016	$[-0.612, 0.579]$	433
High concern	Danger	0.431	$[-0.588, 1.527]$	115
	Correction	-0.195	$[-1.306, 0.781]$	110

Estimates are differences-in-means for wave 2 responses among units that were able to be re-contacted.
Observations weighted by CCES sample weights.
95% percentile confidence intervals obtained from 1000 bootstrap resamples.
$* = p < .05, \dagger = p < .1$.

Table 6: Nyhan and Reifler (2015) – Naive SACE estimates – Wave 2

Subgroup	Treatment	Non-response Wave 2	95% CI	N
Full sample	Danger	-0.006	$[-0.083, 0.079]$	659
	Correction	0.003	$[-0.076, 0.094]$	662
Low concern	Danger	0.044	$[-0.045, 0.145]$	508
	Correction	0.019	$[-0.067, 0.097]$	515
High concern	Danger	-0.157^{*}	$[-0.312, -0.017]$	151
	Correction	-0.025	$[-0.233, 0.176]$	147

Observations weighted by CCES sample weights. Sample sizes are the sum of units in the relevant treatment condition and in the no-message control.
95% percentile confidence intervals obtained from 1000 bootstrap resamples.
$* = p < .05, \dagger = p < .1$.

Table 7: Nyhan and Reifler (2015) – Estimated treatment effects on Wave 2 attrition

is not clear that the treatment substantially affected respondents' propensity to be recontacted in the second wave. That is, the naive Wave 2 effects are not necessarily *biased* causal estimates. Table 7 shows the estimated effects of each treatment on units' average probability of dropping out. Only one sub-group, the high-concern respondents, showed statistically significant evidence

of differential attrition rates across the "danger" treatment. These respondents were actually more likely to respond to the follow-up survey when assigned to the danger condition than if they received no treatment. Of course, as before, with multiple sub-group comparisons, such a significant effect may stukk be due to chance particularly with such a small sample in the "high concern" stratum.

Whether these treatment effects tell researchers anything about attrition rates, however, requires some additional thought about what assumptions are reasonable. Without assuming monotonicity, the absence of an average effect on survival is insufficient enough to show that the non-always/never survivor strata are minimal. Indeed, a zero effect is consistent with no always-survivors and an equal number of protected/harmed respondents. Theory and prior beliefs in this case strongly point to minimal if not zero treatment effects on attrition. Bcause respondents answer many different questions while taking the CCES, the effect of any particular survey survey exposure on follow-up will likely be minimal. Rather, successful follow-up is more likely to be driven by demographic characteristics of the individuals rather than attributes of the survey. For example, respondents with generally more leisure time may be more likely to be available for subsequent re-interviewing.

Table 8 presents the differences in the marginal distributions of five pre-treatment covariates between the full-sample and the subset that responded to the second wave. Re-contacted respondents tend to be better-educated, older, whiter, and less concerned about flu vaccine side-effects. The last variable is a particularly strong source of effect heterogeneity as suggested by the wave 1 analysis. But in general, all of these variables could be expected to be potential sources of variation in treatment effects. Therefore, evaluating whether the difference in wave 1 and wave 2 effects of the correction treatment is due to the treamtent itself requires an effect in wave 2 to be matched to a comparable sample.

The absence of strong evidence for differential attrition, at least for the "concern" treatment, suggests that I can use a more efficient estimator of the re-weighted SACE under the assumption of no individual treatment effect on survival. With that stronger assumption, no principal score weighting is necessary to identify the SACE. Additionally, for the re-weighted SACE, $P(S_i(1) = S_i(0) = 1|X_i)$ can be estimated using a model for observed survival with both treated and control units (rather than just control), reducing the overall variance of the estimator.

Variable	Value	Full Sample	Re-contacted Sample
Sex	Male	0.482	0.488
	Female	0.518	0.512
Age	18 - 29	0.212	0.192
	30-44	0.244	0.247
	45-59	0.282	0.283
	60+	0.262	0.277
Education	High school or less	0.396	0.377
	Some college	0.34	0.336
	College graduate	0.174	0.187
	Post-graduate	0.09	0.1
Race/ethnicity	White	0.712	0.776
	Black	0.123	0.083
	Hispanic	0.102	0.07
	Other	0.063	0.071
Side-effect concern	Extremely concerned	0.111	0.098
	Very concerned	0.125	0.125
	Somewhat concerned	0.318	0.292
	Not too concerned	0.309	0.337
	Not at all concerned	0.136	0.146
Sample Size		1000	882

Observations weighted by CCES sample weights.

Table 8: Nyhan and Reifler (2015) – Covariate distributions – Full sample v. Recontacted sample

Subgroup	Treatment	Vaccine gives flu (1-4)	95% CI	N	N Survival
Full sample	Correction	-0.251	$[-0.550, 0.068]$	660	540
Low concern	Correction	-0.136	$[-0.451, 0.212]$	513	432
High concern	Correction	-0.408	$[-0.890, 0.145]$	147	108

Subgroup	Treatment	Flu vaccine unsafe (1-4)	95% CI	N	N Survival
Full sample	Correction	0.153	$[-0.098, 0.383]$	660	541
Low concern	Correction	-0.019	$[-0.262, 0.199]$	513	432
High concern	Correction	0.431	$[-0.106, 0.947]$	147	109

Subgroup	Treatment	Intend to vaccinate (1-6)	95% CI	N	N Survival
Full sample	Correction	-0.153	$[-0.761, 0.436]$	660	542
Low concern	Correction	-0.114	$[-0.828, 0.605]$	660	541
High concern	Correction	-0.584	$[-1.659, 0.565]$	147	110

Observations weighted by CCES sample weights and stratum selection weights estimated using an additive logistic regression model predicting observed survival with sex, age category, education, race/ethnicity and vaccine side-effect concern. N denotes the number of observations in the full-sample while N_S denotes the number of observations that survived to follow-up
95% percentile confidence intervals obtained from 1000 bootstrap resamples.
$* = p < .05, \dagger = p < .1$.

Table 9: Nyhan and Reifler (2015) – Re-weighted SACE for Correction – No treatment effect on survival assumption

As the original paper's focus is the effect of the corrective treatment condition, and because almost none of the effects of this treatment from wave 1 appear to be significant in the second wave, I apply the re-weighted SACE estimator to assess whether this difference is attributable to variation in sample composition and respondent effect heterogeneity. Table 9 presents estimates for the re-weighted SACE in wave 2 under a "no effect of treatment on survival assumption." I estimate the weights using a logistic regression model fit to predict the observed indicator of survival with the five observed covariates.

I find moderate differences in point estimates for the "full sample" re-weighted wave 2 effects from the unweighted estimates in table 6. These differences are consistent with the sample selection story. Effects from the high concern group were under-represented in the overall treatment effect when averaging over the observations conditional on survival. Whereas the estimated effect of treatment on belief that the vaccine causes the flu using the naive estimator was -0.195, the weighting estimator brings that estimate up to -0.251 (with a p-value just barely above .1).

More interestingly, sample composition does not appear to account for the existence of the "backfire" effect for the high concern group of respondents. While the weighting estimator has higher variance than the naive unweighted one (as the weights have to be estimated as well), re-weighting only negligibly changes the estimated positive effect of the correction on belief that vaccines are unsafe. However, the consequence of higher variance drives the significance level above .1. Results for intent to vaccinate remain statistically insignificant though the point estimates themselves shift sharply negative, suggesting, again, that many of the respondents that reacted negatively to the correction in terms of their stated vaccination preferences were also more likely to be missing in the follow-up sample, thereby attenuating the treatment effect.

Overall, the results suggest that researchers should certainly be concerned about sample attrition as heterogeneity in sub-group effects can substantially change the interpretation of treatment effect estimates even when there is no bias in estimation of the causal quantity. Changes in treatment effects over time may be a result of changes in effective sample composition rather than meaningful short versus long term effects. In the context of the Nyhan and Reifler (2015) study, the degree of non-random attrition is enough to meaningfully alter point estimates of treatment effects. However, it does not appear to account for all of the differences between wave 1 and wave 2, suggesting some differences in the way treatment affects instantaneous versus long-term

responses, particularly in the context of auxiliary beliefs about vaccine safety. While the results are not strongly indicative of a longer-term backfire effect within the high-concern sub-group, sample attrition does not explain every difference between waves one and two.

Conclusion

This paper addresses an issue that many applied researchers struggle with – what to do when subjects drop out of the study and are otherwise unable to have elicited outcomes. In such a situation, meaningful effects are defined only for the units that would not drop out under either treatment or control – the principal stratum (Frangakis and Rubin, 2002) of always-survivors. Unfortunately, identification of these effects is more difficult than it is for an average treatment effect or a conditional treatment effect as selecting on units that do not drop out may induce a form of post-treatment bias, even when the treatment is fully randomized. I outline identification and estimation strategies for such causal effects using ignorability assumptions for survival conditional on some set of measured background covariates, drawing on similar work in the instrumental variables literature.

I also contribute a new result to the literature on survivor effect estimation. Specifically, I show that attrition induces not only internal validity problems, but also external validity problems. Even when researchers can estimate the treatment effect among those units whose survival is unaffected by treatment, and even when treatment *has* no effect on survival, treatment effect estimates can be misleading. This is particularly true when researchers have some baseline estimate to which they want to compare the survivor effect, as is the case in multi-wave studies where the difference in first and second period effects is of theoretical importance. Unadjusted comparisons between full-sample and survivor causal effects are uninformative in the presence of effect heterogeneity. When the set of units for which a causal effect is defined changes, as it does under attrition, differences in effects may be equally attributable to changes in the underlying causal process or simply to heterogeneity of treatment effects among sub-groups and differences in representation of these sub-groups in the survivor group.

The replication of Nyhan and Reifler (2015) shows the utility of analyzing potentially contaminated follow-up data in experiments where non-random sample attrition occurs. Rather than

throw away data that was collected, researchers may want to consider treating attrition in experiments as a secondary observational study. When many covariates are measured, the assumptions underlying identification under principal ignorability become much more plausible. And when treatment is highly unlikely to affect survival, conditioning on survival does not induce bias for the causal estimand. Therefore, researchers do not have biased effects, but they may have *unrepresentative* effects. This representation problem can be corrected by the drawing on weighting tools to extrapolate from the local treatment effect to an effect defined over a more representative set of co-variates, as this paper does.

Of course the major challenge in applying these re-weighting methods is justifying the assumptions needed for identification of the SACE. Monotonicity of treatment effects on survival . In effect, survivor bias converts an experimental design into an observational design, and identification requires choosing one of two fundamentally untestable assumptions – the first being no individual treatment effects on survival, and the second being ignorability of principal stratum assignment conditional on covariates. The former concerns itself with fundamentally unidentifiable individual treatment effects (Holland, 1986) while the latter cannot be guaranteed by randomization. Future work in this area should extend sensitivity analysis methods to principal score estimation of the SACE and allow researchers to implement an easy-to-interpret evaluation of the principal ignorability assumption in addition to existing sensitivity methods for monotonicity. Likewise, when researchers have many covariates to measure, model selection tools for the principal score model become incredibly important as researchers must find some way of obtaining dimensionality reduction with minimal bias in modeling.

References

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of econometrics* 113(2):231–263.

Angrist, Joshua D, Guido W Imbens and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91(434):444–455.

Angrist, Joshua and Ivan Fernandez-Val. 2010. Extrapolate-ing: External validity and overidentification in the late framework. Technical report National Bureau of Economic Research.

Aronow, Peter M and Allison Carnegie. 2013. "Beyond LATE: Estimation of the average treatment effect with an instrumental variable." *Political Analysis* pp. 492–506.

Blackwell, Matthew. 2014. "A selection bias approach to sensitivity analysis for causal effects." *Political Analysis* 22(2):169–182.

Boyd, Christina L, Lee Epstein and Andrew D Martin. 2010. "Untangling the causal effects of sex on judging." *American journal of political science* 54(2):389–411.

Cole, Stephen R and Constantine E Frangakis. 2009. "The consistency statement in causal inference: a definition or an assumption?" *Epidemiology* 20(1):3–5.

Ding, Peng and Jiannan Lu. 2016. "Principal stratification analysis using principal scores." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Ding, Peng, Zhi Geng, Wei Yan and Xiao-Hua Zhou. 2011. "Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death." *Journal of the American Statistical Association* 106(496):1578–1591.

Feller, Avi, Fabrizia Mealli and Luke Miratrix. 2016. "Principal Score Methods: Assumptions and Extensions." *arXiv preprint arXiv:1606.02682* .

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

Hernan, Miguel A and James M Robins. 2017. *Causal inference*. CRC Boca Raton, FL:.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81(396):945–960.

Imai, Kosuke. 2008. "Sharp bounds on the causal effects in randomized experiments with truncation-by-death." *Statistics & probability letters* 78(2):144–149.

Imbens, Guido W. 2004. "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and statistics* 86(1):4–29.

Jo, Booil and Elizabeth A Stuart. 2009. "On the use of propensity scores in principal causal effect estimation." *Statistics in medicine* 28(23):2857–2875.

Mattei, Alessandra, Fabrizia Mealli and Barbara Pacini. 2014. "Identification of causal effects in the presence of nonignorable missing outcome values." *Biometrics* 70(2):278–288.

Nyhan, Brendan and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32(2):303–330.

Nyhan, Brendan and Jason Reifler. 2015. "Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information." *Vaccine* 33(3):459–464.

Robins, James. 1986. "A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect." *Mathematical Modelling* 7(9):1393–1512.

Robins, James M, Miguel Angel Hernan and Babette Brumback. 2000. "Marginal structural models and causal inference in epidemiology." *Epidemiology* 11(5):550–560.

Rosenbaum, Paul R and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* pp. 41–55.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66(5):688.

Rubin, Donald B. 1990. "Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9."." *Statistical Science* 5(4):472–480.

Rubin, Donald B. 2006. "Causal inference through potential outcomes and principal stratification: application to studies with" censoring" due to death." *Statistical Science* pp. 299–309.

Tchetgen Tchetgen, Eric J. 2014. "Identification and estimation of survivor average causal effects." *Statistics in medicine* 33(21):3601–3628.

Tchetgen Tchetgen, Eric J, M Maria Glymour, Ilya Shpitser and Jennifer Weuve. 2012. "Rejoinder: To Weight or Not to Weight?: On the Relation Between Inverse-probability Weighting and Principal Stratification for Truncation by Death." *Epidemiology* 23(1):132–137.

Volfovsky, Alexander, Edoardo M Airoldi and Donald B Rubin. 2015. "Causal inference for ordinal outcomes." *arXiv preprint arXiv:1501.01234* .

Zhang, Junni L and Donald B Rubin. 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.

Zhang, Junni L, Donald B Rubin and Fabrizia Mealli. 2009. "Likelihood-based analysis of causal effects of job-training programs using principal stratification." *Journal of the American Statistical Association* 104(485):166–176.

Appendix A: Proofs of Theorems and Propositions

Proof of Proposition 1 Using assumption 1 for both Y_i and S_i

$$E[\widehat{SACE}] = E[Y_i(1)|A_i = 1, S_i(1) = 1] - E[Y_i(0)|A_i = 0, S_i(0) = 1]$$

$$E[\widehat{SACE}] = E[Y_i(1)|A_i = 1, S_i(1) = 1] - E[Y_i(0)|A_i = 0, S_i(0) = 1]$$

Assumption 1

$$E[\widehat{SACE}] = E[Y_i(1)|S_i(1) = 1] - E[Y_i(0)|S_i(0) = 1]$$

Let $\pi_{ab} = Pr(S_i(1) = a, S_i(0) = b)$

Law of total probability yields

$$\begin{aligned} E[\widehat{SACE}] = {} & E[Y_i(1)|S_i(1) = 1, S_i(0) = 1]\frac{\pi_{11}}{\pi_{11} + \pi_{10}} + \\ & E[Y_i(1)|S_i(1) = 1, S_i(0) = 0]\frac{\pi_{10}}{\pi_{11} + \pi_{10}} - \\ & E[Y_i(0)|S_i(1) = 1, S_i(0) = 1]\frac{\pi_{11}}{\pi_{11} + \pi_{01}} - \\ & E[Y_i(0)|S_i(1) = 0, S_i(0) = 1]\frac{\pi_{01}}{\pi_{11} + \pi_{01}} \end{aligned}$$

Subtracting off the true SACE

$$\begin{aligned} E[\widehat{SACE}] - \text{SACE} = {} & E[Y_i(1)|S_i(1) = 1, S_i(0) = 1]\left(\frac{\pi_{11}}{\pi_{11} + \pi_{10}} - 1\right) + \\ & E[Y_i(1)|S_i(1) = 1, S_i(0) = 0]\frac{\pi_{10}}{\pi_{11} + \pi_{10}} - \\ & E[Y_i(0)|S_i(1) = 1, S_i(0) = 1]\left(\frac{\pi_{11}}{\pi_{11} + \pi_{01}} - 1\right) - \\ & E[Y_i(0)|S_i(1) = 0, S_i(0) = 1]\frac{\pi_{01}}{\pi_{11} + \pi_{01}} \end{aligned}$$

Simplifying

$$E[\widehat{SACE}] - \text{SACE} = (E[Y_i(1)|S_i(1)=1, S_i(0)=0] - E[Y_i(1)|S_i(1)=1, S_i(0)=1])\,\frac{\pi_{10}}{\pi_{11}+\pi_{10}} - (E[Y_i(0)|S_i(1)=0, S_i(0)=1] - E[Y_i(0)|S_i(1)=1, S_i(0)=1])\,\frac{\pi_{01}}{\pi_{11}+\pi_{01}}$$

Proof for Proposition 2 Consistency, positivity and conditional ignorability given X_i

$$\begin{aligned}
E[S_i(a)] = Pr(S_i(a)=1) &= \sum_{x\in\mathcal{X}} Pr(S_i(a)=1|X_i=x)Pr(X_i=x) \\
&= \sum_{x\in\mathcal{X}} Pr(S_i=1|A_i=a, X_i=x)Pr(X_i=x) \\
&= \sum_{x\in\mathcal{X}} Pr(S_i=1|A_i=a, X_i=x)\frac{Pr(A_i=a)}{Pr(A_i=a|X_i=x)}Pr(X_i=x|A_i=a) \\
&= \sum_{x\in\mathcal{X}} Pr(S_i=1|A_i=a, X_i=x)\frac{Pr(A_i=a)}{Pr(A_i=a|X_i=x)}Pr(X_i=x|A_i=a))
\end{aligned}$$

Note that when treatment is fully randomized, $E[S_i(a)] = E[S_i|A_i=a]$.

Substituting in sample-analogues and summing over the empirical distribution yields a consistent estimator of $\widehat{Pr(S_i(a))}$

$$\widehat{Pr(S_i(a))} = \frac{1}{\sum_{i:A_i=a}\frac{1}{\hat{\pi}_i^a}}\sum_{i:A_i=a}\frac{1}{\hat{\pi}_i^a}S_i$$

where $i : A_i = a$ denotes the sum over all units with treatment a and $\frac{1}{\hat{\pi}_i^a}$ is the estimated inverse propensity of treatment weight for observation i.

We can define the principal strata proportions in terms of the two marginal probabilities $Pr(S_i(1)=1)$ and $Pr(S_i(0)=1)$. Under monotonicity, $S_i(0)=1$ implies $S_i(1)=1$. Therefore,

$$\begin{aligned}
Pr(S_i(1)=1, S_i(0)=1) &= Pr(S_i(1)=1|S_i(0)=1)Pr(S_i(0)=1) \\
&= Pr(S_i(0)=1)
\end{aligned}$$

And for the protected stratum

$$\begin{aligned} Pr(S_i(1) = 1, S_i(0) = 0) &= Pr(S_i(1) = 1) - Pr(S_i(1) = 1, S_i(0) = 1) \\ &= Pr(S_i(1) = 1) - Pr(S_i(1) = 0) \end{aligned}$$

Finally, the never-survivor proportions are

$$\begin{aligned} Pr(S_i(1) = 0, S_i(0) = 0) &= 1 - (Pr(S_i|A_i = 1) - Pr(S_i|A_i = 0)) - Pr(S_i|A_i = 0) \\ &= 1 - Pr(S_i(1) = 1) \end{aligned}$$

Proof for Proposition 3 First, consider ignorability conditional on covariates

$$\begin{aligned} E[X_i|S_i(1) = S_i(0) = 1] &= \sum_{x\in\mathcal{X}} xPr(X_i = x|S_i(1) = S_i(0) = 1) \\ &= \sum_{x\in\mathcal{X}} x\frac{Pr(A_i = 0)}{Pr(A_i = 0|X_i = x)}Pr(X_i = x|A_i = 0, S_i(1) = S_i(0) = 1) \\ &= \sum_{x\in\mathcal{X}} x\frac{Pr(S_i(1) = S_i(0) = 1|X_i = x, A_i = 0, S_i = 1)}{Pr(S_i(1) = S_i(0) = 1|A_i = 0, S_i = 1)}\frac{Pr(A_i = 0)}{Pr(A_i = 0|X_i = x)}Pr(X_i = x|A_i = 0, S_i = 1) \\ &= \sum_{x\in\mathcal{X}} x\frac{Pr(A_i = 1)}{Pr(A_i = 1|X_i = x)}Pr(X_i = x|A_i = 0, S_i = 1) \end{aligned}$$

where the last step follows from monotonicity.

Therefore

$$E[X_i|S_i(1) = S_i(0) = 1] = \frac{E[\gamma(x)X_i]}{E[\gamma(x)]}$$

where

$$\gamma(x) = \frac{(S_i(1 - A_i) \cdot (1 - P(A_i = 1))}{1 - P(A_i = 1|X_i)}$$

Proof for Proposition 5 For identifying $E[Y_i(a)|S_i(1) = S_i(0) = 1]$, first use law of iterated expectations with respect to X_i

$$E[Y_i(a)|S_i(1) = S_i(0) = 1] = \sum_{x=\mathcal{X}} E[Y_i(a)|S_i(1) = S_i(0) = 1, X_i = x]Pr(X_i = x|S_i(1) = S_i(0) = 1)$$

Ignorability of treatment assignment

$$\sum_{x=\mathcal{X}} E[Y_i(a)|A_i = a, S_i(1) = S_i(0) = 1, X_i = x]Pr(X_i = x|S_i(1) = S_i(0) = 1)$$

And principal ignorability

$$\sum_{x=\mathcal{X}} E[Y_i(a)|A_i = a, S_i(a) = 1, X_i = x]Pr(X_i = x|S_i(1) = S_i(0) = 1)$$

Consistency gives

$$\sum_{x=\mathcal{X}} E[Y_i|A_i = a, S_i = 1, X_i = x]Pr(X_i = x|S_i(1) = S_i(0) = 1)$$

Bayes' Rule:

$$\sum_{x=\mathcal{X}} E[Y_i|A_i = 1, S_i = 1, X_i = x]Pr(X_i = x|A_i = a, S_i(1) = S_i(0) = 1)\frac{Pr(A_i = a|S_i(1) = 1, S_i(0) = 1)}{Pr(A_i = a|X_i = x, S_i(a) = 1, S_i(0) = 1)}$$

Ignorability of treatment for survival potential outcomes.

$$\sum_{x\in\mathcal{X}} E[Y_i|A_i = 1, S_i = 1, X_i = x]\frac{Pr(A_i = a)}{Pr(A_i = a|X_i = x)}Pr(X_i = x|A_i = a, S_i(1) = S_i(0) = 1)$$

Consistency for survival

$$\sum_{x\in\mathcal{X}} E[Y_i|A_i = 1, S_i = 1, X_i = x]\frac{Pr(A_i = a)}{Pr(A_i = a|X_i = x)}Pr(X_i = x|A_i = a, S_i = 1, S_i(1-a) = 1)$$

By Bayes' rule

$$\sum_{x\in\mathcal{X}} E[Y_i|A_i = a, S_i = 1, X_i = x]\frac{Pr(S_i(1-a) = 1|X_i = x, A_i = a, S_i = 1)}{Pr(S_i(1-a) = 1|A_i = a, S_i = 1)}\frac{Pr(A_i = a)}{Pr(A_i = a|X_i = x)}Pr(X_i = x|A_i = a, S_i = 1)$$

Ignorability of treatment for survival potential outcomes.

$$\sum_{x\in\mathcal{X}} E[Y_i|A_i=a,S_i=1,X_i=x]\frac{Pr(S_i(1-a)|X_i=x,S_i(a)=1)}{Pr(S_i(1-a)=1|,S_i(a)=1)}\frac{Pr(A_i=a)}{Pr(A_i=a|X_i=x)}Pr(X_i=x|A_i=a,S_i=1)$$

Definition of conditional probability

$$\sum_{x\in\mathcal{X}} E[Y_i|A_i=a,S_i=1,X_i=x]\frac{\frac{Pr(S_i(1)=1,S_i(0)=1|X_i=x)}{Pr(S_i(a)=1|X_i=x)}}{\frac{Pr(S_i(1)=1,S_i(0)=1)}{Pr(S_i(a)=1)}}\frac{Pr(A_i=a)}{Pr(A_i=a|X_i=x)}Pr(X_i=x|A_i=a,S_i=1)$$

Marginal distribution of $S_i(a)$ identified by ignorability

$$\sum_{x\in\mathcal{X}} E[Y_i|A_i=a,S_i=1,X_i=x]\frac{\frac{Pr(S_i(1)=1,S_i(0)=1|X_i=x)}{Pr(S_i=1|A_i=1,X_i=x)}}{\frac{Pr(S_i(1)=1,S_i(0)=1)}{Pr(S_i=1|A_i=a)}}\frac{Pr(A_i=a)}{Pr(A_i=a|X_i=x)}Pr(X_i=x|A_i=a,S_i=1)$$

Monotonicity identifies the stratum proportions

$$\sum_{x\in\mathcal{X}} E[Y_i|A_i=a,S_i=1,X_i=x]\frac{\frac{Pr(S_i=1,|A_i=0,X_i=x)}{Pr(S_i=1|A_i=a,X_i=x)}}{\frac{Pr(S_i=1|A_i=0)}{Pr(S_i=1|A_i=a)}}\frac{Pr(A_i=a)}{Pr(A_i=a|X_i=x)}Pr(X_i=x|A_i=a,S_i=1)$$

Which yields an expression in terms of observable quantities

$$\frac{Pr(A_i=a)Pr(S_i=1|A_i=a)}{Pr(S_i=1|A_i=0)}\sum_{x\in\mathcal{X}} E[Y_i|A_i=a,S_i=1,X_i=x]\frac{Pr(S_i=1,|A_i=0,X_i=x)}{Pr(S_i=1|A_i=a,X_i=x)}\frac{1}{Pr(A_i=a|X_i=x)}Pr(X_i=x|A_i=a,S_i=1)$$

Substituting in the sample analogues and summing over the empirical distribution of X yields the estimator

$$E[Y_i(a)|\widehat{S_i(1)=S_i(0)=1}] = \frac{1}{\sum_{i:A_i=a,S_i=1}\frac{\hat{w_i^a}}{\hat{\pi_i^a}}}\sum_{i:A_i=a,S_i=1}\frac{\hat{w_i^a}Y_i}{\hat{\pi_i^a}}$$

where $i : A_i = a, S_i = 1$ denotes all units i with treatment a and survival status 1, $\frac{1}{\hat{\pi_i^a}} = \frac{\widehat{Pr(A_i=a)}}{\widehat{Pr(A_i=a|X_i=x)}}$ denotes the stabilized inverse propensity weight for the estimated probability that unit i receives treatment a and $\widehat{w_i^a} = \frac{\widehat{Pr(S_i=1|A_i=0,X_i=x)}}{\widehat{Pr(S_i=1|A_i=a,X_i=x)}}$ is the estimated principal score conditional on survival. Note that for $a = 0$, this weight is 1 (as monotonicity implies all units with $S_i = 1, A_i = 0$ are always-survivors).

Assuming the models for the estimated weights are consistent and correctly specified for their population analogues, $E[Y_i(a)|\widehat{S_i(1)=S_i(0)=1}]$ is consistent for the true $E[Y_i(a)|S_i(1) = S_i(0) = 1]$.

Therefore

$$\widehat{SACE}^{PS} = \frac{1}{\sum_{i:A_i=1,S_i=1} \frac{\hat{w_i^1}}{\hat{\pi_i^1}}} \sum_{i:A_i=1,S_i=1} \frac{\hat{w_i^1} Y_i}{\hat{\pi_i^1}} - \frac{1}{\sum_{i:A_i=0,S_i=1} \frac{1}{\hat{\pi_i^0}}} \sum_{i:A_i=a,S_i=1} \frac{Y_i}{\hat{\pi_i^0}}$$

is consistent for the true SACE.

Proof for Proposition 6 The goal is to estimate the conditional SACE averaged over the full-sample distribution of X.

$$\sum_{x \in \mathcal{X}} E[Y_i(1) - Y_i(0) | S_i(1) = S_i(0) = 1, X_i = x] Pr(X_i = x)$$

By Bayes' Rule

$$\sum_{x \in \mathcal{X}} E[Y_i(1) - Y_i(0) | S_i(1) = S_i(0) = 1, X_i = x] Pr(X_i = x | S_i(1) = 1, S_i(0) = 1) \frac{Pr(S_i(1) = S_i(0) = 1)}{Pr(S_i(1) = S_i(0) = 1 | X_i = x)}$$

From Proposition 2 we can write the principal strata probabilities in terms of observables

$$\sum_{x \in \mathcal{X}} E[Y_i(1) - Y_i(0) | S_i(1) = S_i(0) = 1, X_i = x] Pr(X_i = x | S_i(1) = 1, S_i(0) = 1) \frac{Pr(S_i = 1 | A_i = 0)}{Pr(S_i = 1 | A_i = 0, X_i = x)}$$

Let $\hat{q_i} = \frac{\widehat{Pr(S_i=1|A_i=0)}}{\widehat{Pr(S_i=1|A_i=0,X_i=x)}}$ be the estimated inverse principal stratum selection weight.
Then, applying the result from Proposition 5, we get that

$$\widehat{RWSACE}_t^{PS} = \frac{1}{\sum_{i:A_i=1,S_i=1} \frac{\hat{q_i}\hat{w_i^1}}{\hat{\pi_i^1}}} \sum_{i:A_i=1,S_i=1} \frac{\hat{q_i}\hat{w_i^1} Y_i}{\hat{\pi_i^1}} - \frac{1}{\sum_{i:A_i=0,S_i=1} \frac{\hat{q_i}}{\hat{\pi_i^0}}} \sum_{i:A_i=a,S_i=1} \frac{\hat{q_i} Y_i}{\hat{\pi_i^0}}$$

is consistent for the re-weighted SACE in a population with a covariate distribution matching the overall sample.