Biometrics 63, 641–662
September 2007

DOI: 10.1111/j.1541-0420.2007.00847.x

Principal Stratification Designs to Estimate Input Data Missing Due to Death

Constantine E. Frangakis,[1,*] Donald B. Rubin,[2] Ming-Wen An,[1] and Ellen MacKenzie[3]

[1]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.
[2]Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.
[3]Department of Health Policy and Management, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.
[*]*email:* cfrangak@jhsph.edu

Summary. We consider studies of cohorts of individuals after a critical event, such as an injury, with the following characteristics. First, the studies are designed to measure "input" variables, which describe the period before the critical event, and to characterize the distribution of the input variables in the cohort. Second, the studies are designed to measure "output" variables, primarily mortality after the critical event, and to characterize the predictive (conditional) distribution of mortality given the input variables in the cohort. Such studies often possess the complication that the input data are missing for those who die shortly after the critical event because the data collection takes place after the event. Standard methods of dealing with the missing inputs, such as imputation or weighting methods based on an assumption of ignorable missingness, are known to be generally invalid when the missingness of inputs is nonignorable, that is, when the distribution of the inputs is different between those who die and those who live. To address this issue, we propose a novel design that obtains and uses information on an additional key variable—a treatment or externally controlled variable, which if set at its "effective" level, could have prevented the death of those who died. We show that the new design can be used to draw valid inferences for the marginal distribution of inputs in the entire cohort, and for the conditional distribution of mortality given the inputs, also in the entire cohort, even under nonignorable missingness. The crucial framework that we use is principal stratification based on the potential outcomes, here mortality under both levels of treatment. We also show using illustrative preliminary injury data that our approach can reveal results that are more reasonable than the results of standard methods, in relatively dramatic ways. Thus, our approach suggests that the routine collection of data on variables that could be used as possible treatments in such studies of inputs and mortality should become common.

Key words: Causal inference; Censoring by death; Missing data; Potential outcomes; Principal stratification; Quantum mechanics.

1. Introduction

We consider studies that interview cohorts of individuals after a critical event, such as injury or stroke, with the following two characteristics. First, the studies are designed to measure "input" variables, which describe the period before the critical event, and to characterize the distribution of the input variables in the cohort. Second, the studies are designed to measure "output" variables, primarily mortality after the critical event, and to characterize the predictive (or conditional) distribution of mortality given the input variables in the cohort. Such studies, however, are often complicated by the fact that the input data are missing for those who die shortly after the critical event because the data are collected after the event.

This problem, input data missing due to death, occurs commonly, for example, in studies of elders (Cornoni et al., 1993; Reuben et al., 1995; Cohen et al., 2002), or victims of injuries (e.g., MacKenzie et al., 2006). The goals we address for such studies are how to estimate the inputs missing due to death, and how to characterize the predictive (or conditional) distribution of mortality given the input variables in the cohort. Answers to these goals are important because, first, they can be used to better alert the individuals and their physicians about increases in risks, and second, they inform about the pathways of such risks.

As a motivating example, consider the National Study on the Costs and Outcomes of Trauma Centers (NSCOT; MacKenzie et al., 2006). That study used hospital discharge records to identify and enroll individuals who received care for injuries. The first follow-up visit was scheduled at 3 months. During this visit, patients were interviewed about their preinjury disability, as measured by "activities of daily living (ADL)." It is of interest to evaluate the relation that prior disability has to the risk of death following an injury. However, some patients died as a result of injury, before this first follow-up visit. Thus, the ADL values are missing for these patients. If these missing past ADL values have a different distribution than the observed past ADL values among survivors, standard methods cannot estimate that relation.

Table 1
Examples of studies with input data missing due to death

Population; original goal	Measures of interest (time 0)	Critical event (time 1)
Elders or sick; relate functional measures to mortality	ADL, intense emotional stress, intense physical activity	Stroke, falls, myocardial infarction, opportunistic infections
Youths; relate exposure measures to severe injury/mortality	Controlled substance use (e.g., alcohol, drug abuse)	Injuries (e.g., crash)

Another class of examples arises in the evaluation of the effect that a periodic exposure (e.g., to drug) has on the risk of a critical event using a case-crossover design (Maclure, 1991). In its basic form, this design aims to measure, for each one of a group of injury cases, the gap time between the last exposure and the critical event, and a measure of that person's typical frequency of past exposure. A measure of association between exposure and the critical event is then defined by comparing the observed gap times to their distribution that would be expected if the critical event had been unrelated to the exposure process defined by the past frequencies. In this design, even if we know the victims' most recent exposure to drugs (e.g., by blood measurement), the frequency of past exposure becomes missing for those who die as a result of severe injuries, and this missingness is usually ignored (e.g., Vinson et al., 1995). As discussed below, such missingness needs to be addressed by new and more appropriate methods. Such examples are summarized in Table 1.

Standard methods confronted with missing data from death, as also noted by Zhang and Rubin (2003), can be classified into three types. The first type is concerned only with the observed data (e.g., cause-specific hazards, dating to Prentice et al., 1978; and partly conditional on being alive, Kurland and Heagerty, 2005); these methods are not relevant to our problem because they do not attempt to estimate the missing data. The second type of method assumes ignorability (Rubin, 1976) of missing data and essentially replaces them with data matched from fully observed strata, either across time from the same person, or across people for the same time (e.g., McMahon and Harrell, 2001; Lin, McCullouch, and Mayne, 2002) or both; these methods are known to be inappropriate when the distribution of data missing due to death differs from that in observed strata (Rubin, 1978). The third type posits nonignorable assumptions relying simply on the parametric structure of models (e.g., Fairclough, Peterson, and Chang, 1998); these methods are sensitive to the parametric assumptions because, without such assumptions, the distributions of interest are not identifiable unless additional design structure is introduced.

We address the problem's goals from a combination of design and analyses perspectives. First, we recognize that the problem is related to, but differs from, the problem of censoring by death discussed in Rubin (2000), Frangakis and Rubin (2002), and developed by Zhang and Rubin (2003). The goal of the latter problem is to compare treatments on potential outcomes (Neyman, 1923; Rubin, 1974, 1978) when some patients in either treatment die. In that problem, the *future* outcome of a person who dies is "missing," not because it exists and is unobserved, but because it is not defined. Because the patients who die may not be comparable between the two treatments, death creates the need to define meaningful treatment effects on the outcomes. Such effects are well defined if we restrict attention to patients who would survive no matter which treatment they would receive (Rubin, 2000) rather than to the larger group of patients who are observed to survive. This group of patients, who would survive no matter the treatment, is a special case of a "principal stratum" (Frangakis and Rubin, 2002), that is, here, a stratum defined by a patient's joint potential outcomes of death under the two treatments. Thus, in that case, the principal strata are critical for defining treatment effects. In the present problem, the variable of interest is a well-defined input preceding death, and is missing because the attempt to record it takes place after death. The key, from the design perspective, then, is to recognize that the missing data of an individual who dies would be observed "under explicit alternative conditions for which the same individual would have survived." Formalizing this, we show that it is also important here for the goal of estimating the missing information, that: (1) the *design* finds data on factors (e.g., treatments) that (1a) could have prevented deaths and (1b) were assigned to the individuals after the time when the inputs of interest became defined but before the time of death; and (2) these data be analyzed using principal stratification.

In the next section, we formulate more explicitly the problem and its goals, and formalize the proposed design with data on externally controllable factors, such as treatments, that can prevent deaths. In Section 3, we describe a method that can address our goals using the data from the proposed design and the framework of principal stratification. We show that the proposed method allows the distribution of missing inputs to differ systematically from the distribution of the observed inputs, yet this method is able to estimate the distribution of the missing inputs. In Section 4, we demonstrate using preliminary data from NSCOT including transport time to hospital as the externally controllable factor, that our design and analysis method can uncover results that are dramatically different and more plausible than those of standard methods. Section 5 provides extensions of the proposed methods in more general situations. Section 6 discusses the commonalities and differences between this and other related uses of principal stratification. Section 7 concludes with remarks, including connections between this new, interventional approach to missing data and the principles of quantum mechanics.

2. Design Using Principal Stratification

2.1 *Initial Design and Goals*

Consider a cohort of individuals who had a critical event (at time say $t = 1$), such as an injury (e.g., crash). We are interested in learning about a variable A that takes its value at a time, say $t = 0$, before the critical event and so is called an input. For example, A can be ADL that the person cannot perform, or exposure to drugs. To record A, we schedule an interview at a time, say $t = 2$, after the critical event, e.g., an interview at discharge from the hospital. However, a subset of individuals die before the interview, as a result of the critical event; for those individuals, the value of A still exists, because it occurred before death, but becomes missing because there is no interview.

Throughout, we use i to index an individual. Let A_i be the value of A for individuals at $t = 0$; and let $S_i^{\text{obs}} = 1$ for surviving individuals at $t = 2$, and 0 otherwise. This initial setting is shown in Figure 1(a).

Goals. We wish to address the following: (a) Estimate the distribution of the past input A_i for the people who died without reporting them; and (b) Estimate quantities such as predictive distributions and associations that are defined based on the distribution of all values A_i, missing and observed, for example, the prediction of death based on A_i. The first goal is important for characterizing the distribution of the inputs for all individuals. The second goal differs from predicting death from the *observed* inputs in this study, $P(S_i^{\text{obs}} = 0 \mid \{A_i : A_i \text{ is observed}\})$, which is by definition deterministically 0 and is of no interest. Goals of type (b) are important because they inform us about the degree to which the past inputs A_i in the original cohort are actually related to death (or to the critical event using additional data from people without that event). Because of the deaths, the inputs A are not all reported *in this* study, so these relations need to be estimated indirectly. These relations should suggest better monitoring methods in *subsequent* studies, which would alert physicians and individuals about sudden increases in the risk of death. Also, goals (b) contribute by helping medical research understand the pathways through which those inputs relate to critical events and death.

2.2 *New Design Elements and Principal Strata*

Consider the following additional design elements:

(i) For all individuals, we find and record a factor or treatment (labeled Z_i) that was assigned externally (i.e., by a person or process other than the individual), and a level of which could have prevented death for those who died. For this factor, let $z = 0$ denote a standard level, and $z = 1$ denote the more effective level. For example, for injuries, such a "treatment factor" can be the *transport time* (long or short) from the time of injury to arrival at the hospital or to surgery, whereas for strokes or myocardial infarctions, such a factor can be the prompt administration of a thrombolytic drug.

(ii) We also record covariates X_i that were used to decide the level Z_i of the factor for the individual. The variables X_i may correlate with the input A_i.

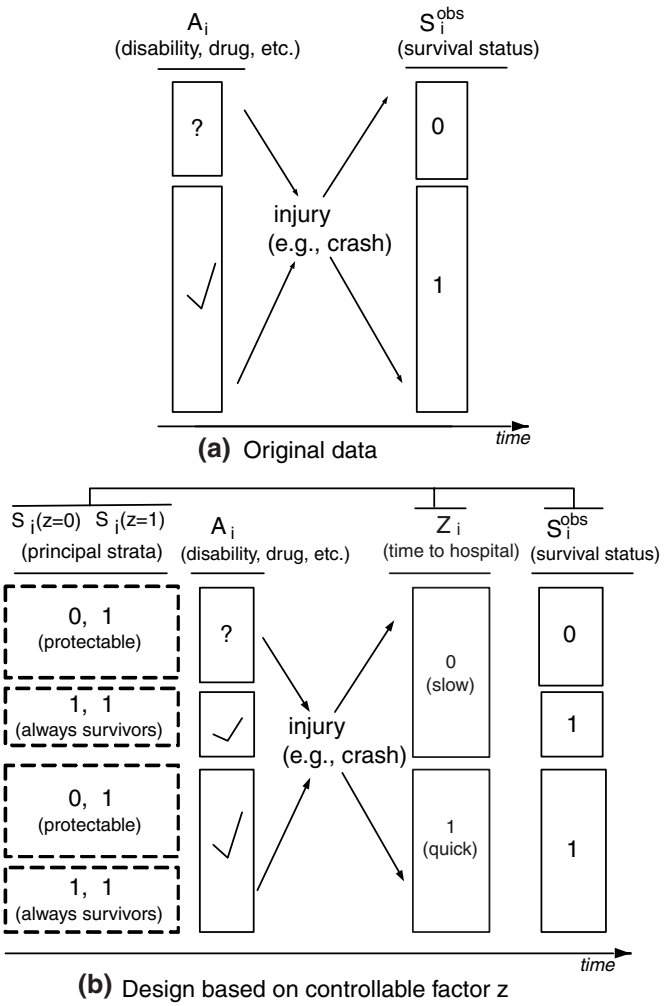

Figure 1. (a): Initial design on input variable A and survival status S^{obs}, matched for past covariates; (b): New design based on a controllable factor. Dashed boxes indicate principal strata with respect to survival. The presentational order from left to right of (principal strata $(S_i(0), S_i(1))$ and input A_i), controllable factor Z_i, and observed survival S_i^{obs}, which determines measurement or no measurement of the input A_i, is also the time order of definition from earliest to latest variable. Other covariates defined before the controllable factor can be used as in Section 5.2.

The level of factor z to which a particular individual is assigned can affect the future of that individual, although we assume it cannot affect the future of a different individual (no interference, Rubin, 1978; Cox, 1992). For an individual i, denote by $S_i(z)$ the potential survival outcome (Rubin, 1978) that indicates the survival status if the individual is assigned level z of the factor. It is then important, as in Rubin (2000), Frangakis and Rubin (2002), and Zhang and Rubin (2003), to consider the principal strata of survival, that is, the strata of the individuals with respect to the joint values of $(S_i(0), S_i(1))$. These are generally the following: (1) individuals who would survive no matter the level of z, that is, $S_i(0) = S_i(1) = 1$; (2) individuals who would die under the standard

level but would live under the effective one, that is, $S_i(0) = 0$ and $S_i(1) = 1$; (3) individuals who would die no matter the level, that is, $S_i(0) = S_i(1) = 0$; and (4) individuals who would survive under the standard level but would die under the effective treatment, that is, $S_i(0) = 1$ and $S_i(1) = 0$. We denote the principal stratum of individual i by P_i and label the above four possible strata as "always survivors," "protectable," "never survivors," and "defiers," respectively, combining terminology of Angrist, Imbens, and Rubin (1996), and Gilbert, Bosch, and Hudgens (2003) for vaccines.

Our main argument is that addressing the goals (a) and (b) can be helped by recording and using data on such a factor z (there can be more than one choice) that can justify plausible assumptions about the assignment of Z_i and about the principal strata.

A simple example reveals how our structure can help us achieve our goals. Consider a factor z that can justify the following two assumptions (for extensions see Section 5):

ASSUMPTION 1: *Ignorable assignment of external factor: The levels Z_i are independent of (A_i, P_i) conditionally on the variables X_i that were used for assignment.*

ASSUMPTION 2: *Preventability of deaths from external factor: Individuals are either P_i = "protectable" by the effective level $(z = 1)$ of the factor, or else "always survivors."*

Assumption 1 is plausible when we choose z and X_i so that conditionally on X_i the reasons for the remaining variability of Z_i are independent of the individuals' health prior to the critical event. For example, we can ask physicians to tell us all the variables they used to decide assignment of a treatment z. So, the external assignment of z makes its ignorability achievable, whereas this is not true for an assumption of "ignorability of death," which is typically made by the standard methods (Section 1). Note that, by definition, the values of A_i and P_i are not affected by the actual treatment that is assigned (Frangakis and Rubin, 2002). The second assumption excludes "never survivor" and "defier" patients, and is related to the monotonicity assumption in other settings (e.g., Angrist et al., 1996). Preventability, when combined with ignorability, is testable from the observed data, because under these assumptions we must observe that among individuals within levels of X_i and assigned the "effective" treatment, all survive, whereas among those assigned the standard treatment some die and some survive, as in Figure 1(b). More generally, some notion of both, the ignorability of the controllable factor, and a type of monotonic effect of that factor on the reason of missing outcomes (here mortality) are critical for using this design. Nevertheless, the preventability assumption is more flexible than it originally appears when made within levels of the covariate strata X_i. The preventability Assumption 2 can also be relaxed to allow for "never survivors" as discussed in Section 5.1. We now show how the above design addresses our goals.

3. Estimability of Input Data Missing Due to Death

3.1 *Distribution of Missing Input Data*

For the observed data, we assume without loss of generality that we are already within covariate strata $X_i = x$; so, for brevity we omit the explicit conditioning on X_i in the notation of the distributions below. The possibly missing input A_i is taken as an indicator for poor functional ability (e.g., dichotomized ADL = 1 for poor status).

Consider first the goal of estimating the distribution of the missing functional inputs, $P(A_i = 1 \mid S_i^{\text{obs}} = 0, Z_i = 0)$. The above ignorability of the assignment of the prevention factor levels Z_i reflects that, conditionally on the variables X_i, and on which we have already stratified, the assignment of Z_i balances all other covariates, including the input A_i, which is a covariate that took its value before the prevention factor Z_i was assigned, even though assignment of Z_i preceded the time when A_i was to be measured. In other words, because A_i is a covariate and Z_i is effectively randomized (given X_i), the proportion $P(A_i = 1 \mid Z_i = 0)$ of poor inputs among individuals assigned the standard prevention level of z equals the proportion $P(A_i = 1 \mid Z_i = 1)$ among those assigned the effective prevention level. Because the former group includes both individuals with observed and missing values, we have that:

$$\begin{aligned} &P(A_i = 1 \mid Z_i = 1) \\ &\quad = P(A_i = 1 \mid Z_i = 0) \\ &\quad = \sum_{s=0,1} P\big(A_i = 1 \,\big|\, S_i^{\text{obs}} = s, Z_i = 0\big) P\big(S_i^{\text{obs}} = s \,\big|\, Z_i = 0\big). \end{aligned} \tag{1}$$

From the observed data, as Figure 1(b) shows, we can estimate directly the proportion $P(A_i = 1 \mid Z_i = 1)$ of people who had had poor function among those assigned the effective level of z. The equality in (1) then implies that we can also estimate the proportion $P(A_i = 1 \mid Z_i = 0)$ of people who had had poor function among those assigned the standard level of z. Moreover, Figure 1(b) shows that we can also directly estimate from the observed data: the proportion $P(S_i^{\text{obs}} = 1 \mid Z_i = 0)$ of survivors among individuals assigned the standard z; and the proportion $P(A_i = 1 \mid S_i^{\text{obs}} = 1, Z_i = 0)$ who had poor function among those who survived after being assigned the standard level of factor z. It follows then, from (1), that the distribution of missing past inputs can be expressed as

$$P\big(A_i = \,|\, S_i^{\text{obs}} = 0, Z_i = 0\big) = \frac{P\big(A_i = 1 \,|\, Z_i = 1\big) - P\big(A_i = 1 \,|\, S_i^{\text{obs}} = 1, Z_i = 0\big) P\big(S_i^{\text{obs}} = 1 \,|\, Z_i = 0\big)}{P\big(S_i^{\text{obs}} = 0 \,|\, Z_i = 0\big)}. \tag{2}$$

Therefore, we have reduced the unknown distribution of missing input data to an expression, the right-hand side of (2), that involves quantities that can be directly estimated as discussed above. This calculation is related to the instrumental variables' equations of the effect of a treatment on

posttreatment outcomes in a trial with noncompliance (Imbens and Rubin, 1997). However, the context and goal of the problem here are different, and this parallel arises from the more fundamental commonality of "principal stratification" shared between the two types of problems (see Section 6).

3.2 *Relation Between Input and Mortality*

The ability to estimate better the missing data allows us to also examine better relations between those data and clinical variables. As an example, we show here how we can estimate the degree to which the input A_i predicts death. Because death depends on the principal strata P_i and the level of the prevention factor, it is important to examine if the input A_i predicts the principal strata of death. This would indicate that A_i predicts the underlying predisposition of a person to die.

Specifically, we wish to estimate:

$$P\{S_i(0) = 0 \mid A_i = a\} = \frac{P\{S_i(0) = 0\}P\{A_i = a \mid S_i(0) = 0\}}{P(A_i = a)}, \tag{3}$$

and compare (3) with $a = 0$ and 1. From the top of (1), we have that $P(A_i = 1)$ equals the directly estimable proportion $P(A_i = 1 \mid Z_i = 1)$ under the effective prevention level. Moreover, from ignorability of treatment assignment with respect to the principal strata, we have that the protectable patients $\{i : S_i(0) = 0\}$ are balanced between the levels of z (all probabilities are implicitly given X_i), and so $P\{S_i(0) = 0\}$ in the right-hand side of (3) equals the directly estimable proportion $P(S_i^{\text{obs}} = 0 \mid Z_i = 0)$ of patients who die under the standard prevention level, where the principal strata are observed (see Figure 1(b)). Also by ignorability, the proportion $P\{A_i = a \mid S_i(0) = 0\}$ of protectable patients who have input a, involved in the right-hand side of (3), is also balanced between the levels of z and so equals the proportion of patients with input a among those who die in the standard prevention level, i.e., $P(A_i = a \mid S_i^{\text{obs}} = 0,\ Z_i = 0)$, where the latter is estimable from (2). These arguments show estimability of the proportions in (3). Using these arguments to substitute the right-hand side of (3) with estimable quantities based on (2), we can express the relative risk of being a protectable (not always survivor) patient when having poor versus good input A_i as

$$\frac{P\{S_i(0) = 0 \mid A_i = 1\}}{P\{S_i(0) = 0 \mid A_i = 0\}} = \frac{P(A_i = 0 \mid Z_i = 1)}{P(A_i = 1 \mid Z_i = 1)} \times \frac{P(A_i = 1 \mid Z_i = 1) - P\big(A_i = 1 \,\big|\, S_i^{\text{obs}} = 1, Z_i = 0\big)P\big(S_i^{\text{obs}} = 1 \,\big|\, Z_i = 0\big)}{P\big(S_i^{\text{obs}} = 0 \,\big|\, Z_i = 0\big) - P(A_i = 1 \mid Z_i = 1) + P\big(A_i = 1 \,\big|\, S_i^{\text{obs}} = 1, Z_i = 0\big)P\big(S_i^{\text{obs}} = 1 \,\big|\, Z_i = 0\big)}, \tag{4}$$

where the quantities in the right-hand side of equation (4) are all directly estimable as described in the paragraph following (1).

4. Demonstration

We return to the NSCOT study (MacKenzie et al., 2006) on injuries described in Section 1. To illustrate the contrast between our approach to missing data and standard approaches, we consider patients who have sustained injuries with a relatively low ($X_i = 0$) or high ($X_i = 1$) severity ($n = 354$, 135 respectively). The follow-up interview is scheduled 3 months after the injury to measure by questionnaire the functional status ($A_i = 1$ for poor ADL) that existed before injury, and this is missing if the injured person i dies before the interview as a result of the injuries. The prevention factor z we use here is based on the time it took to transport the injured person to the hospital.

Regarding the assumption of ignorability of the assignment mechanism of the transport time to hospital, the two main reasons for variability of this time are (a) the severity of the injury as judged by medical personnel—more severe injuries are attempted to be transported faster; and (b) external reasons such as time of day, distance, traffic, or weather, that prevent fast transport, but that are themselves in principle not directly related to the person's health before injury. It is therefore plausible to assume ignorable assignment of Z_i after conditioning on the measured severity of injury X_i used to decide Z_i: among individuals of the same injury severity X_i (high or low, see Table 2), those transported slowly are assumed to have the same distributions of past ADL A_i and principal strata P_i as the individuals transported quickly. Of course, one may wish to adjust for additional levels of covariates to remove possible remaining confounding, for example, using the approach of Section 5.2, but the principles for those analyses remain the same. The assumption that quick transportation to hospital can prevent an important proportion of deaths is supported both by literature for other critical events (e.g., GISSI, 1986), and empirically by our data: within either of our strata (high or low) of injury severity X_i, there were no deaths for injuries delivered to the hospital within 10 minutes, although there were 19% deaths for patients with a high injury severity delivered later than 10 minutes and 5% deaths for patients with a low injury severity delivered later than 10 minutes. Based on the above, Table 2 gives relevant summary proportions, directly computed from the data. We treat these summaries here as population proportions, because they indicate plausible results for each method. Inferential statements were not planned for and so do not achieve statistical significance, because the study had not been planned to use the new design.

Focusing first on high injury severity, there were $P(Z_i = 1) = 8\%$ of patients transported quickly; among the patients who were transported slowly, 81% survived, i.e., $P(S_i^{\text{obs}} = 1 \mid Z_i = 0) = 81\%$; among those transported quickly, all survived, i.e., $P(S_i^{\text{obs}} = 1 \mid Z_i = 1) = 100\%$ (not shown); of those, there were 9% who had poor ADL before injury, i.e., $P(A_i = 1 \mid Z_i = 1) = 9\%$; and among those who survived after being transported slowly, 5% had poor past A_i, i.e., $P(A_i = 1 \mid S_i^{\text{obs}} = 1,\ Z_i = 0) = 5\%$. Then, the approach that would estimate the protectable patients' missing data distribution

Table 2
Demonstration of design using injury data from NSCOT

Data					Results based on new design and method	
Injury severity X_i	% transported quickly $P(Z_i = 1)$	Survival (%) when transported slowly $P(S_i^{obs} = 1 \mid Z_i = 0)$	Input ADL when transported quickly $P(A_i = 1 \mid Z_i = 1)$	Input ADL of survivors when transported slowly $P(A_i = 1 \mid S_i^{obs} = 1, Z_i = 0)$	Input ADL of those who died when transported slowly $P(A_i = 1 \mid S_i^{obs} = 0, Z_i = 0)$	Relative risk of death as in expr (4)
Low	2%	95%	25%	22%	82%	13.7
High	8%	81%	9%	5%	26%	3.6
				Standard methods would assume these the "same"		Standard methods would assume these = 1

$P(A_i = 1 \mid S_i^{obs} = 0, Z_i = 0)$ with the distribution of observed data after matching on slow time $Z_i = 0$ would give 5% poor function. On the other hand, an approach that would estimate the missing data distribution with the observed data without matching on time would give $P(A_i = 1 \mid S_i^{obs} = 1)$ which equals $\sum_z P(A_i = 1 \mid S_i^{obs} = 1, Z_i = z) \frac{P(S_i^{obs}=1|Z_i=z)P(Z_i=z)}{\sum_{z'} P(S_i^{obs}=1|Z_i=z')P(Z_i=z')}$, and which, using the information given in Table 2, gives 5.4%. More generally, the result of the standard methods is bounded to be between the directly observed $P(A_i = 1 \mid S_i^{obs} = 1, Z_i = z)$, for $z = 0, 1$ (here, between 5% and 9%), as a convex combination of the two.

With the new method, however, the missing proportion of poor past function for protectable patients is allowed to be different from the observed strata. In particular, from (1), the missing proportion $P(A_i = 1 \mid S_i^{obs} = 0, Z_i = 0)$ must be such that when mixed with the proportion of $P(A_i = 1 \mid S_i^{obs} = 1, Z_i = 0) = 5\%$ of poor past function for always survivors, the result should be the proportion of $P(A_i = 1 \mid Z_i = 0) = P(A_i = 1 \mid Z_i = 1) = 9\%$ observed for all patients transported quickly to the hospital (Figure 1(b)). The fact that, by (1), this is a convex mixing based on the probabilities $P(S_i^{obs} = s \mid Z_i = 0)$, for $s = 0, 1$, implies that the missing proportion $P(A_i = 1 \mid S_i^{obs} = 0, Z_i = 0)$ of poor past function for the protectable patients must be *higher* than the mixture, $P(A_i = 1 \mid Z_i = 1) = 9\%$. Using (2), the missing proportion $P(A_i = 1 \mid S_i^{obs} = 0, Z_i = 0)$ is $\{9\% - (5\%)(81\%)\}/(100\% - 81\%) = 26\%$. This shows that the actual result can be estimable and substantially different from those of the standard methods. Note that this proportion is in line with a hypothesis that those who died had generally poorer past ADL than the survivors. Analogous comparisons are obtained for injuries with low severity. Finally, the larger proportions of poor ADL for low versus high injury severity is in accordance with the hypothesis that individuals who sustain injuries of light severity *and* who, nevertheless, *need* hospitalization, were more frail before the injury than individuals who get hospitalized after sustaining a severe injury.

The relative risk in (4) is implicitly assumed to equal 1 by the standard method that replaces the missing data distribution $P(A_i = 1 \mid S_i^{obs} = 0, Z_i = 0)$ with that of the observed data after matching on the prevention level, that is, with $P(A_i = 1 \mid S_i^{obs} = 1, Z_i = 0)$. With the new method, however, and the empirical proportions of Table 2, the relative risk in (4) is estimated to be 13.7 and 3.6, for low and high injury severity, respectively. This means that, even after conditioning on observed strata, the possibly missing functional ability is an important predictor of the underlying ability of a patient to survive the injury when transportation takes a standard time to the hospital. The first implication is that follow-up, e.g., of individuals with history of poor functionality, should use new designs (e.g., based on automated reporting devices) to make sure that some dimensions of functional ability be measured at higher frequency. This would give better prediction for which patients transition to high risk for death from a critical event. The second implication is that sudden changes to low functional ability inputs should be examined medically to understand and address the pathways through which these inputs predict death from injury even in the short term.

5. More General Role of New Methods

5.1 *Partial Preventability*

The new methods are important also for more general input data, designs, and assumptions. A plausible prevention factor may partly, but not fully, prevent death. For example, prompt delivery of thrombolytic drugs prevents death after stroke in some but not all cases (GISSI, 1986). More specifically for such settings, we consider an external factor z that satisfies no interference and Assumption 1, as in Section 2.2, and a generalization of Assumption 2:

ASSUMPTION 3: *Partial preventability of deaths from external factor: Individuals are either P_i = "never survivors," "protectable," or "always survivors."*

For never survivors—those who would not survive no matter the factor's level—the observation of outcomes then remains essentially undefined just based on this factor, and so is not estimable without further assumptions. So the goal in this setting is limited to the estimation of the distribution of missing inputs for protectable patients under

the standard level of assignment, which equals $P(A_i \mid P_i =$ protectable) by Assumption 1. Standard methods cannot estimate correctly this distribution, as they cannot do so in the setting given in Sections 2 and 3. Yet we show below that this distribution is still estimable without further assumptions.

To see this, note that the distribution of observed inputs under the effective factor level, as in Section 2.2, is still a mixture of the distribution among protectables and always survivors. Letting p, a stand for protectables and always survivors, respectively, we then have

$$\begin{aligned} &P\big(A_i = 1 \,\big|\, S_i^{\text{obs}} = 1, Z_i = 1\big) \\ &\quad = \sum_{q=p,a} P\big(A_i = 1 \,\big|\, P_i = q, S_i^{\text{obs}} = 1, Z_i = 1\big) \\ &\qquad \times P\big(P_i = q \,\big|\, S_i^{\text{obs}} = 1, Z_i = 1\big) \\ &\quad = \sum_{q=p,a} P(A_i = 1 \mid P_i = q) \times P(P_i = q)/P(P_i \in \{p, a\}), \end{aligned} \tag{5}$$

where the last equality for the first summand arises first, because S^{obs} is a function of P and Z, and then because A, P is independent of Z, by Assumption 1.

To recover the target of interest, $P(A_i = 1 \mid P_i = p)$, from (5), note that, because among the patients assigned the effective level $Z_i = 1$, those who survive are the protectables and always survivors, the proportions $P(S_i^{\text{obs}} = 1 \mid Z_i = 1)$ and $P(P_i \in \{p,\ a\})$ are equal. Moreover, because among those assigned the standard level, $Z_i = 0$, those who survive are always survivors, it follows that the proportions $P(S_i^{\text{obs}} = 1 \mid Z_i = 0)$ and $P(P_i = a)$ are equal, and the distribution of input data $P(A_i = 1 \mid P_i = a)$ equals the directly estimable distribution $P(A_i = 1 \mid S^{\text{obs}} = 1, Z_i = 0)$. By substituting these in (5) and after some rearrangement of terms we find that the target distribution satisfies

$$\begin{aligned} &P(A_i = 1 \mid P_i = p) \\ &\quad = \frac{P\big(A_i = 1 \,|\, S_i^{\text{obs}} = 1, Z_i = 1\big) P\big(S_i^{\text{obs}} = 1 \,|\, Z_i = 1\big) - P\big(A_i = 1 \,|\, S_i^{\text{obs}} = 1, Z_i = 0\big) P\big(S_i^{\text{obs}} = 1 \,|\, Z_i = 0\big)}{P\big(S_i^{\text{obs}} = 1 \,|\, Z_i = 1\big) - P\big(S_i^{\text{obs}} = 1 \,|\, Z_i = 0\big)}, \end{aligned}$$

The last expression, although similar to (2) for full preventability, is different, first, in the left side of the numerator, which now also measures the likelihood of staying alive after assignment to the effective level $Z_i = 1$, and in the denominator, which, instead of $P(S^{\text{obs}} = 0 \,|\, Z_i = 0)$, now expresses the proportion of protectables in the case of partial preventability.

The above result means that for the subset of patients that are protectable or always survivors we can still assess the ignorability of missingess of data, and also find the direction along which its violation occurs (e.g., if such input data for those who died were higher on average than the observed ones). Thus in such more general settings, the importance of the new methods is essentially intact for addressing the scientific goals.

5.2 *Modeling Covariates*

Suppose we still make Assumptions 1 and 3, but we first wish to condition on multiple, and possibly continuous, covariates X_i, and that to do so, we model the distribution of the principal strata of survival and of a continuous input given principal strata by parametric functions

$$\begin{aligned} l^{(\text{P})}\big(q, x, \beta^{(\text{P})}\big) &:= P\big(P_i = q \mid X_i = x, \beta^{(\text{P})}\big), \quad \text{and} \\ l^{(\text{A})}(a, q, x, \beta^{(\text{A})}) &:= P\big(A_i = a \mid P_i = q, X_i = x, \beta^{(\text{A})}\big), \end{aligned} \tag{6}$$

where the last function is defined only for $q =$ protectable, or always survivor. Denote by $\mathcal{P}(Z_i, S_i^{\text{obs}})$ the set of possible principal strata as a function of the observed level Z_i and survival status S_i^{obs}: if $Z_i = 0$ (standard) and $S^{\text{obs}} = 1$ (alive), then $\mathcal{P}(Z_i, A_i^{\text{obs}}) = \{$always survivor$\}$; if $Z_i = 0$ and $S^{\text{obs}} = 0$ (dead), then $\mathcal{P}(Z_i, S_i^{\text{obs}}) = \{$protectable, never survivor$\}$, if $Z_i = 1$ (effective) and $S^{\text{obs}} = 0$ (dead), then $\mathcal{P}(Z_i, S_i^{\text{obs}}) = \{$never survivor$\}$, and if $Z_i = 1$ and $S^{\text{obs}} = 1$ (alive), then $\mathcal{P}(Z_i, S_i^{\text{obs}}) = \{$protectable, always survivor$\}$. Then the likelihood of the collection of data

$$X_i, Z_i, S_i^{\text{obs}}, \quad \text{and} \quad A_i \text{ if } S_i^{\text{obs}} = 1,$$

over independent individuals, conditional on the covariates and the observed factor levels, is

$$\begin{aligned} &\text{Likd}\big(\beta^{(\text{P})}, \beta^{(\text{A})}\big) \\ &\quad = \prod_i \sum_{q \in \mathcal{P}(Z_i, S_i^{\text{obs}})} l^{(\text{P})}\big(q, X_i, \beta^{(\text{P})}\big) \cdot \big\{ l^{(\text{A})}\big(A_i, q, X_i, \beta^{(\text{A})}\big) \big\}^{S_i^{\text{obs}}}. \end{aligned} \tag{7}$$

Under this setting, we can more generally express a quantity of interest as a function $Q(\beta^{(\text{P})}, \beta^{(\text{A})})$ of the parameters, which can then be estimated by using likelihood or Bayesian methods to estimate the parameters from (7). Semiparametric methods, as discussed by Scharfstein, Rotnitzky, and Robins (1999) in general, and by Gilbert et al. (2003) for an application of principal stratification to vaccine trials, are also of interest. The fact that these quantities would be identifiable by our method even without the models in (6) if samples were large enough means that the results should not be sensitive to the particular parametric models, as long as they are flexible. Moreover, we can also show better estimation of general quantities of importance in Table 2, such as for associations using case-crossover designs.

6. Related Problems

The design and structure of principal stratification we proposed for this problem, "inputs missing due to death," has commonalities and also differences with the structure of two other problems where studies assign a treatment to examine its effect on an outcome. The first problem, "treatment

noncompliance," deals with subjects who do not comply with the assigned treatment, and its structure with principal strata was discussed by Imbens and Rubin (1997). The second problem, "outcomes censored by death" (see Section 1), deals with subjects who die *before* the intended *future* outcome is measured, and its structure with principal strata has been discussed by Rubin (2000), Frangakis and Rubin (2002), Zhang and Rubin (2003), and, with adaptation to HIV vaccines, by Gilbert et al. (2003).

The common structure across these problems is centered around a factor that can be thought of as controllable, in the sense that its assignment is assumed ignorable. All three problems also have factors whose values are measurable after the controllable factor is assigned, namely postcontrollable (or endogenous) factors; and factors whose values are defined (but not necessarily measurable) before the controllable factor is assigned, namely precontrollable factors. The latter include all *potential* outcomes of the postcontrollable factors, and, therefore, include principal strata, that is, crossclassifications of subjects by some subset of potential outcomes. The three problems also have differences, in their structure, their goals, and in the role that principal stratification plays in addressing these goals.

In the problem with "treatment noncompliance," the controllable factor is the treatment assignment; the postcontrollable factors are the observed treatment received and the outcome; and the precontrollable factors are the potential values of the treatment received and of the outcome. Of particular importance is the principal stratum of "compliers," that is, the subjects for whom the potential values of treatment received are the same as the treatment assigned, for all assignment levels (Imbens and Rubin, 1997). In this problem, the principal stratification helps to formulate and, under assumptions, estimate the effect of treatment assignment (or intention to treat) on the outcome among the compliers. This goal is important because for compliers, the experimental comparison of outcomes among the levels of the controlled assignment is also a comparison among the different levels of treatment received.

In the problem with "outcomes censored by death," the controllable factor is again the treatment assignment; the postcontrollable factors are the observed survival status, and the observed outcome if the person survives; and the precontrollable factors are the potential values of the survival and of the outcome. Here, a principal stratum of particular importance is that of "always survivors," defined as in Section 2. Principal stratification helps formulate and estimate the effect of treatment assignment on the outcome among always survivors. This goal is important because always survivors are the only subjects for whom potential outcomes are well defined for all assignment levels.

In the problem with "inputs missing due to death," the controllable factor is one that affects survival after a critical event, that is, *an event after which, under standard conditions, there is substantial likelihood of death, such as injury or stroke*; the postcontrollable factors are the observed survival status of the person, which determines measurement (if alive) or no measurement (if dead) of the input that occurred before the critical event; and the precontrollable factors are the inputs of interest and the principal strata of survival *that take their value before the measurement of the critical event*. Here, the "protectables" are a principal stratum of particular importance. In this problem, the principal stratification provides the framework for appropriately positing assumptions, such as those of Sections 2, 3, or 5, that allow estimation of the distribution of the missing inputs for protectable patients. This goal is important because it better characterizes the differences between observed and missing inputs, and helps better understand the role that the inputs have for predicting mortality from the critical event.

7. Discussion

We proposed a framework for addressing data missing due to death by obtaining and using data and explicit assumptions about a treatment assignment mechanism that could cause missing values to become observed if different levels of the treatment had been assigned. Thus, although a relation between causal inference and missing data has been obvious since Neyman (1923) and Rubin (1974, 1976, 1978), the proposed framework for data missing due to death emphasizes a particular order for understanding these concepts: causal inference with potential outcomes is not just a special case of missing data, but is *more* fundamental than missing data (see also Rubin, 1987, 2005). Specifically, in the framework we proposed, data can only be regarded as having a missing value if an explicit intervention can be proposed that would provide us with that value. This principle for missing data, therefore, follows the principle of quantum mechanics, by which a measurable value of a physical quantity is only defined in terms of an explicit intervention that can be applied in order to provide that value. This parallel of principles is also reflected in the parallel of primary elements of the two frameworks—the complex wave function in quantum mechanics, and the principal strata of potential outcomes in the proposed framework for missing data: these primary elements give rise to the observed data by specific rules, but the primary elements are not themselves directly observable, providing an additional dimension that empowers the frameworks to better explain observations.

The use of an intervention factor to address missing data has the limitation that there can be settings where such a factor can exist, but still not be available in the design. This can be so especially because such factors are not, at present systematically recorded for the purposes of addressing missing data, because their role in this problem had not previously been demonstrated. For such cases where the missing values are well defined but where design features do not allow their identifiability, sensitivity analyses can be implemented (e.g., Rubin, 1977; Manski, 2003). Our results and illustration, though, demonstrate that using such intervention factors can improve the evaluation of and utility of studies with missing data due to death, and so can be the first step to a more systematic recording of such factors.

It will also be of interest to combine the setting discussed here, where possible deaths of patients can imply that their unobserved past is different from pasts that are observed, with the settings considered by Rubin (2000) and Zhang and Rubin (2003). In those settings, patients who die could have

had also a different *future* outcome trajectory from observed trajectories, under conditions that would have prevented their death. Developing methods to answer such combined questions is important for evaluating, for example, not only the potential benefit of prevention programs for saving lives, but also the programs' effects on the quality of patients' lives, and the relation of these effects to past input variables.

Acknowledgements

We thank the editor, the associate editor and a reviewer for helpful comments, Dan Scharfstein for discussions, and the projects on "Statistical methods for partially controlled studies" (NEI), "Center for prevention and early intervention" (NIMH), "Disparities Among Children Served by the CMHS Children Service's Program" (NIA), and the "NSCOT study" (NCIPC, CDC) for partial support.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with Discussion). *Journal of the American Statistical Association* **91,** 444–472.

Cohen, H. J., et al. (2002). A controlled trial of inpatient and outpatient geriatric evaluation and management. *New England Journal of Medicine* **346,** 905–912.

Cornoni-Huntley, J., et al. (1993). Established populations for epidemiologic studies of the elderly: Study design and methodology. *Aging (Milano)* **5,** 27–37.

Cox, D. R. (1992). Causality: Some statistical aspects. *Journal of the Royal Statistical Society, Series A* **155**, part 2, 291–301.

Fairclough, D. L., Peterson, H. F., and Chang, V. (1998) Why are missing quality of life data a problem in clinical trials of cancer therapy? *Statistics in Medicine* **17,** 667–677.

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

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in AIDS vaccine trials. *Biometrics* **59,** 531–541.

GISSI. (1986). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico. *The Lancet* **1**(8478), 397–402.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* **25,** 305–327.

Kurland, B. F. and Heagerty, P. J. (2005). Directly parametrized regression conditioning on being alive: Analysis of longitudinal data truncated by deaths. *Biostatistics* **6,** 241–258.

Lin, H., McCulloch, C. E., and Mayne, S. T. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine* **21,** 2369–2382.

MacKenzie, E. J., Rivara, F. P., Jurkovich, G. J., Nathens, A. B., Frey, K. P., Egleston, B. L., Salkever, D. S., and Scharfstein, D. O. (2006). A national evaluation of the effect of trauma-center care of mortality. *New England Journal of Medicine* **354,** 366–378.

Maclure, M. (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* **133,** 144–153.

Manski, C. F. (2003). *Partial Identification of Probability Distributions.* New York: Springer.

McMahon, R. P. and Harrell, F. E. Jr. (2001). Joint testing of mortality and a non-fatal outcome in clinical trials. *Statistics in Medicine* **20,** 1165–1172.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. Translated in *Statistical Science* **5,** 465–480, 1990.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34,** 541–554.

Reuben, D., Borok, G., Wolde-Tsadik, G., Ershoff, D., Fishman, L., Ambrosini, V., Liu, Y., Rubenstein, L., and Beck, J. (1995). Randomized trial of comprehensive geriatric assessment in the care of hospitalized patients. *New England Journal of Medicine* **332,** 1345–1350.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Rubin, D. B. (1977). Formalizing subjective notions about the effects of nonrespondents in sample surveys. *Journal of the American Statistical Association* **72,** 538–543.

Rubin, D. B. (1978). Bayesian inference for causal effects. *Annals of Statistics* **6,** 34–58.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D. B. (2000). Comment on "Causal inference without counterfactuals," by A. P. Dawid. *Journal of the American Statistical Association* **95,** 435–437.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **469,** 322–331.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94,** 1096–1146.

Vinson, D. C., Mabe, N., Leonard, L. L., Alexander, J., Becker, J., Boyer, J., Moll, J. (1995). Alcohol and injury. A case-crossover study. *Archives of Family Medicine* **4,** 505–511.

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

Received May 2006. *Revised October* 2006.
Accepted November 2006.

Discussions

James Robins,[1,3] Andrea Rotnitzky,[2,3] and Stijn Vansteelandt[4]

[1] *Department of Epidemiology*
Harvard School of Public Health, 655 Huntington Avenue
Boston, Massachusetts 02115, U.S.A.

[2] *Department of Economics*
Universidad Torcuato di Tella
Saenz Valiente 1010
1428 Buenos Aires, Argentina
email: *arotnitzky@utdt.edu*

[3] *Department of Biostatistics*
Harvard School of Public Health, 655 Huntington Avenue
Boston, Massachusetts 02115, U.S.A.

[4] *Department of Applied Mathematics and Computer Sciences*
Ghent University, Krijgslaan 281 S9
9000 Ghent, Belgium

We are grateful for the opportunity to discuss this article. In this discussion, we (i) question the plausibility of the authors' substantive assumptions, (ii) discuss the authors' choice of scientific goals and their attainability, (iii) comment on statistical issues, and (iv) describe a sensitivity analysis approach to the authors' problem.

(i) Substantive Assumptions. In Section 3, the authors show that under no interference and assumptions (I) $(A, S(0), S(1)) \perp\!\!\!\perp Z \,|\, X, C = 1$, where C is the binary indicator of the critical event, e.g., car accident, and (II) $S(1) = 1$ w.p.1, an application of Bayes' Theorem implies identification of the joint distribution $f(A, S \,|\, C = 1, X)$ from the distributions $f(A \,|\, S = 1, C = 1, X)$ and $f(S \,|\, C = 1, X)$.

Assumption 2 stipulates a dichotomous treatment factor Z, which is guaranteed to prevent death. In the authors' example, Z was transport time to hospital, a continuous variable that was dichotomized at 10 minutes. As the authors recognize, treatments Z satisfying (II) rarely exist. For instance, a fraction of individuals injured in car accidents die almost immediately. For them, a "transport time to hospital" of less than 10 minutes cannot prevent death. Yet, the empirical analysis of Section 4 reports that no deaths occurred in subjects with transport times of less than 10 minutes. Possible explanations for the lack of deaths in the rapidly transported would include (i) ambulance paramedics appropriately transport victims found dead at the scene less quickly than injured survivors, (ii) the chosen cutpoint of 10 minutes was data driven, and (iii) the number of high-risk rapidly transported subjects (i.e., 11) was sufficiently small that all survived by chance.

In Section 5, the authors replace assumption (II) with the monotonicity assumption that Z cannot cause death. However, it can be difficult to find variables Z that satisfy monotonicity. For instance, the authors suggest that thrombolytic drug therapy after stroke is a treatment that never causes death. Yet, physicians are well aware that thrombolytic drugs can cause intracerebral hemorrhage and death. Similarly, rapid transport to a hospital may cause death if, in their hurry, the paramedics fail to properly stabilize the patient. Indeed, it is a matter of debate whether fast transport is harmful or beneficial for accident victims (Lerner et al., 2003).

We also question the validity of assumption (I). Subjects with limited preaccident physical mobility (X) both have difficulty with activities of daily living (A) and are difficult to quickly extract from a wrecked automobile. We doubt one could measure physical mobility sufficiently well to insure $A \perp\!\!\!\perp Z \,|\, X, C = 1$ holds, for Z "transport time."

In conclusion, we regard neither the monotonicity assumption (much less the stronger assumption (II)) nor the ignorability assumption (I) as plausible in the authors' examples.

(ii) Scientific goals. In Section 2.1, the authors list their scientific goals as estimation of (a) $f(A \,|\, S = 0, C = 1)$ and (b) $P(S = 0 \,|\, A = a, C = 1)$ as a function of a.

Attainability of the authors' goals. The authors show that $f(A \,|\, S = 0, C = 1)$ and $P(S = 0 \,|\, A = a, C = 1)$ are identified under (I) and (II). In fact, a calculation using Bayes' rule shows that they are identified under the weaker assumptions $A \perp\!\!\!\perp Z \,|\, X, C = 1$ and $\mathrm{P}(S = 1 \,|\, Z = C = 1, X) = 1$. These assumptions do not require any reference to or assumptions about counterfactuals. Unfortunately, the arguments in (i) above show that these weaker assumptions are also unrealistic.

In spite of our concerns about the monotonicity assumption, we now examine whether the authors' goals are

attainable when this assumption holds. In Section 5.1, the authors state that the importance of their approach "is essentially intact for addressing scientific goals," even when (II) is replaced by the weaker monotonicity assumption. We disagree because $f(A \mid S = 0, C = 1)$ and $P(S = 0 \mid A = a, C = 1)$ are not identified, and thus not consistently estimable, when only (I) and monotonicity are imposed.

Perhaps the authors' claim is predicated on the fact that under (I) and monotonicity, the principal stratum distributions $f(A \mid C = 1, P =$ 'Z prevents death') and $f(A \mid C = 1, P =$ 'always survive') are identified. However, the following example demonstrates that knowledge of these distributions does not suffice to address important scientific questions when $f(A \mid S = 0, C = 1)$, and thus $f(A \mid C = 1)$, remain unidentified.

Example. Suppose Z is an antibird-flu drug that is in limited supply and $C = 1$ is contracting bird flu. Clearly, all else being equal, we should give the drugs to those most likely to be helped by the drug. Thus, we would like to know if $P(Z$ prevents death $\mid A = 1, C = 1) > P(Z$ prevents death $\mid A = 0, C = 1)$ for, then, we should give the drug to subjects with $A = 1$ rather than $A = 0$. By Bayes' Theorem, this inequality is $P(A = 1 \mid C = 1, P =$ 'Z prevents death'$)/P(A = 0 \mid C = 1, P =$ 'Z prevents death'$) > P(A = 1 \mid C = 1)/P(A = 0 \mid C = 1)$. When Z does not cause death but is not guaranteed to prevent death, we can identify the left-hand side of the inequality but we cannot identify its right-hand side and thus cannot determine whether the inequality is true.

Relevance of the authors' goals. The preceding example illustrates that knowledge of $f(A \mid S = 0, C = 1)$ can help address substantive questions. However, we argue that $P(S = 0 \mid A = a, C = 1)$ is not relevant for predicting survival when Z is available. If, as the authors assume, data on a strong predictor Z are available, then clearly $P(S = 0 \mid A = a, C = 1, Z = z)$ is a more relevant predictive distribution than $P(S = 0 \mid A = a, C = 1)$. Indeed, if Z were a widely available nontoxic medical treatment that never caused death, it would be unethical to withhold Z and so $P(S = 0 \mid A = a, C = 1, Z = 1)$ would be the only predictive distribution of interest. Note that this implies that obtaining data on Z is a good idea irrespective of whether data on A are missing.

The authors state that knowledge of $P(S = 0 \mid A = a, C = 1)$ (or, when data on X are collected, of $P[S = 0 \mid A = a, C = 1, X]$) helps "medical research understand the pathways through which those inputs relate to critical events and death." The authors did not provide any justification for this claim. Furthermore, they did not define what they meant by the term "pathway." To evaluate the authors' claim we first clarify the meaning of this term. Because our discussion applies even when no data are missing, we may assume A is always observed.

The term "pathways" is generally used as shorthand for "causal pathways." Consider the query: does A have a causal effect on survival S through a pathway that does not involve the critical event C? This query is often rephrased as whether A has a direct causal effect on survival not through C. The concept of direct effect has been formalized in three different ways. Let $S(a)$ and $C(a)$ denote a subject's counterfactual survival and critical event outcome when A is set to a, which we take to be well defined. The subject's observed data S and C are $S = S(A)$ and $C = C(A)$ with A the observed treatment. Let the counterfactual $S(a, c)$ denote a subject's survival when A and C are set to a and c. When $S(a, c)$ is well defined, $S(a)$ equals $S(a, C(a))$. Suppose that, unlike earlier subsections, X is a variable that is causally unaffected by either A or C. The average *controlled direct effect* of A on S when C is set to c within levels of X is defined as $CDE(c) = E[S(1, c) - S(0, c) \mid X]$ (Robins, 1986, 1987). The average *pure direct effect* of A on survival not through C given X is defined as $PDE = E[S(1, C(0)) - S(0, C(0)) \mid X] = E[S(1, C(0)) - S(0) \mid X]$. This contrast measures the average effect of A on survival when C is set to its value $C(0)$ under $A = 0$ (Robins and Greenland, 1992; Pearl, 2001). The *principal stratum average direct effect* of A on survival at level c given X is defined as $PSDE(c) = E[S(1) - S(0) \mid X, C(0) = C(1) = c]$ (Frangakis and Rubin, 2002).

The conditioning subset in $PSDE(c = 1)$ consists of those with covariate X who always suffer the critical event. Robins (1986, Section 12.2) used this contrast to address the problem of censoring by competing causes of death, with $S = 1$ denoting death from a cause of interest (subsequent to a time t) and $C = 0$ denoting death from competing causes (before t). Subsequently, Robins (1995), Rubin (1998, 2000, 2006), Little and Rubin (1999), Robins and Greenland (2000), and Frankagis and Rubin (2002) also employed this contrast in addressing "censoring by death." Baker (2000), Frankagis and Rubin (2002), Gilbert, Bosch, and Hudgens (2003), Rubin (2004), Shepherd et al. (2006), Hudgens and Halloran (2006), and Matsuyama and Morita (2006) used this contrast to address a number of other causal issues.

The contrasts $CDE(c)$ and PDE are well defined only when $S(a, c)$ is well defined. In contrast, $PSDE(c)$ is well defined whenever $S(a)$ and $C(a)$ are well defined. How do we decide whether a counterfactual is well defined? This has been a hotly debated issue in philosophy. The following example, from Quine (1950), effectively ended counterfactual analysis among philosophers until the late 1960s. "If Bizet and Verdi had been of the same nationality, they both would have been French." Quine argued that, because Bizet was French and Verdi Italian, by symmetry considerations, this counterfactual was neither true nor false and thus was ill defined. Lewis (1973) later rejoined that, even though some counterfactuals may be ill defined and all are somewhat vague, many are useful. Robins and Greenland (2000) agreed but went further. They argued that counterfactuals are "vague" to the degree to which one fails to make precise the hypothetical interventions.

Following Robins and Greenland (2000) and Baker (2000), we believe that for subjects with $C = 0$, the intervention corresponding to setting C to 1 is ill defined because (i) $C = 1$ only encodes the occurrence of an accident, failing, for example, to distinguish high-speed head-on collisions from rear-enders at moderate speed and (ii) there is no basis for choosing among them as the intervention. As a consequence $S(a, c)$ is ill defined. Thus among the three direct effects, only $PSDE(c)$ is well defined. Unfortunately, the following somewhat humorous example demonstrates that knowledge of $PSDE(c)$ may add little to our understanding of the pathways by which A relates to critical events and death. Like the authors, we restrict attention to $PSDE(c = 1)$.

Example. Suppose a psychiatrist hypothesizes that, conditional on preaccident health and the seriousness of the crash injury, hypervigilant, controlling individuals ($A = 1$) have higher postaccident in-hospital mortality ($S = 0$) than distractible laid-back individuals ($A = 0$). His theory is that the loss of personal control during hospitalization causes controlling individuals to have serious life-threatening arrhythmias. Suppose intervention on A is well defined. For example, there might exist drugs that can change a person from state $A = 1$ to $A = 0$ (Prozac and Valium) and vice versa (amphetamines). Suppose, conditional on X, $PSDE(c = 1)$ is very negative. Does this refute a skeptic who believes the psychiatrist's hypothesis is false? It does not because $PSDE(c = 1)$ would also be negative under the following scenario. The psychiatrist's hypothesis is false. However, hypervigilant individuals avoid most potential accidents. Those they cannot avoid are usually serious head-on collisions with speeding cars that cross the centerline, leaving no time to react. In contrast, distractible, laid-back individuals have frequent, less serious collisions, because they are neither in a hurry nor do they look where they are going. Thus, individuals in the stratum "always an accident" will tend to have serious accidents and thus a high in-hospital mortality rate when $A = 1$, but less serious accidents when $A = 0$. Thus, a negative $PSDE(c = 1)$ may arise because $A = 1$ increases mortality over $A = 0$; (i*) by directly causing increased in-hospital mortality, as hypothesized by the psychiatrist or, (ii*) solely by preventing minor accidents, as in the last scenario. We conclude that negative values of $PSDE(c = 1)$ fail to indicate the presence of direct effects of A not through its effect on accidents.

The difficulties with $PSDE(c = 1)$ are due to the fact that the event $C = 1$ lumps together the occurrence of accidents of varying severity. Thus, the natural solution is to replace C with a multivariate variable C^* that records relevant details of an accident including the type and seriousness of the injuries sustained. Then a nonzero C^*-specific principal stratum contrast $PSDE(c^*)$ could still be explained by pathway (i*) but no longer by (ii*), thus surmounting the difficulties of $PSDE(c = 1)$. Unfortunately, replacing C with C^* creates a major problem for the principal stratum approach: there is no subject with $C^*(0) = C^*(1)$ if, as is likely, A has an effect on at least one component of every subject's C^*. In that case, the event $C^*(0) = C^*(1) = c^*$ has probability zero for all c^*, rendering the principal stratum approach useless. Even if there were subjects with $C^*(0) = C^*(1)$, their numbers would likely be few. Consequently, the principal stratum approach would only apply to a small subset of the population. Robins and Rotnitzky (2007) catalogue analogous difficulties in substantively important examples. We believe these difficulties are sufficiently problematic to suggest that the principal stratum approach to direct effects is, at times, of little scientific value.

Counterfactuals regained. As we record more details in C^*, the intervention that sets C^* to c^* and the counterfactual $S(a, c^*)$ becomes less and less vague. Consequently, $CDE(c^*)$ and $PDE^* = E[S(1, C^*(0)) - S(0, C^*(0)) \mid X]$ will often be reasonably well defined. In our opinion, these are the contrasts that best serve to distinguish among different pathways. For example, they distinguish pathway (i*) from (ii*) above: PDE^* or $CDE(c^*)$ equal to 0 for all c^* is consistent with (ii*) but not with (i*), while nonzero values of PDE^* or $CDE(c^*)$ can be explained by (i*) but not by (ii*). Of course, even $S(a, c^*)$ is somewhat vague. The only counterfactuals free of vagueness are the treatment-assignment potential outcomes of a randomized experiment, but they are often uninformative about pathways. Because PDE^* only requires $S(a, c^*)$ to be defined for $a = 1$ and $c^* = C^*(0)$, there exist studies in which PDE^* may be regarded as well defined even when $CDE(c^*)$ is not for some C^* (Petersen, Sinisi, and van der Laan, 2006).

We end this section by noting that none of the three contrasts $CDE(c^*)$, PDE^*, and $PSDE(c)$ are identifiable from knowledge of $P(S = 0 \mid A = a, C = 1, X)$, $f(A \mid C = 1, X)$, and $f(X \mid C = 1)$ without additional strong assumptions that were not either assumed or considered by the authors. We conclude that, even had the authors succeeded in their goal of learning these distributions, this success would not have helped "understand the pathways through which inputs relate to critical events and death."

(iii) Statistical issues. In Section 6, the authors discuss similarities between the problem treated in Section 5 and the problem of treatment noncompliance in randomized trials. We now show that these problems are statistically not merely similar but isomorphic. As a consequence, (i) some of the material in Section 5 simply reproves previously published results and (ii) doubly robust semiparametric methods already exist (Tan, 2006) that address the modeling issues of Section 5.2 and do not require that Z be dichotomous.

Assumptions of Section 5.2 are exactly the same as the monotonicity, exclusion, and randomization assumptions considered in the noncompliance literature, upon appropriate identification of the authors' variables with those in a noncompliance model. Specifically, identify X with a prerandomization variable, Z with randomized arm, and $S(z)$ with the actual treatment received when $Z = z$. Then $S = S(Z)$. In the authors' problem, A is a variable that is uninfluenced by Z and would be recorded, if, possibly contrary to fact, the person survived. Thus, we can regard A as the potential outcome $A(s = 1, z) = A(s = 1)$ for any z. This identity is the exclusion restriction. Further, assumption (I) is the assumption that Z is randomized and the assumption that Z never causes death is the monotonicity assumption. Under these assumptions, Abadie (2003) has shown that $E[A \mid S(1) > S(0), X] = \frac{\pi(X,1)\eta(X,1) - \pi(X,0)\eta(X,0)}{\pi(X,1) - \pi(X,0)}$ where $\eta(x, z) = E[A \mid S = 1, X = x, Z = z]$ and $\pi(x, z) = P[S = 1 \mid X = x, Z = z]$. The right-hand side is precisely the right-hand side of the last displayed equation in Section 5.1 in the case of no X's. Tan (2006) showed how to obtain doubly robust estimators of $E[A \mid S(z) > S(z')] = \frac{E[\pi(X,1)\eta(X,1) - \pi(X,0)\eta(X,0)]}{E[\pi(X,1) - \pi(X,0)]}$ when $z > z'$, with high-dimensional X and Z possibly nonbinary, even continuous, that are consistent if either a working model for $f_Z[z \mid X = x]$ is correct or working models for both $\pi(x, z)$ and $\eta(x, z)$ are correct.

(iv) A sensitivity analysis. Because we wish not to impose assumptions (I) and (II), the distributions $f(A \mid S = 0, C = 1)$ and $P(S = 0 \mid A = a, C = 1)$ of interest are

not identified. Instead we suggest a sensitive analysis motivated by the observations that (a) $f(A \mid S = 0, C = 1)$ and $P(S = 0 \mid A = a, C = 1)$ would be identified were $P(A = 1 \mid C = 1)$ identified and (ii) with $W \equiv (X, Z)$, $P(A = 1 \mid C = 1)$ is identified under the nonparametric just-identified nonignorable model for $\pi(W, A) \equiv P(S = 0 \mid W, A, C = 1)$ that specifies $\pi(W, A) = \{1 + \exp\{-[h(W) + Q]\}\}^{-1}$ where $h(W)$ is an unknown function and $Q = q(A, W)$ is a user-specified selection bias function. However, because Q itself is not identified, we later vary it in a sensitivity analysis. Because W is high dimensional, we also specify flexible parametric models $B(\eta) = b(W; \eta)$ and $h(W; \alpha)$ for $b(W) \equiv E[A\exp(Q) \mid C = S = 1, W]/E[\exp(Q) \mid C = S = 1, W]$ and $h(W)$. We compute the estimators $(\hat{\alpha}, \hat{\eta})$ given in Scharfstein, Rotnitzky, and Robins (1999) and $\hat{P}(A = 1|C = 1)$ as the sample average over $C = 1$ of $[S\{1 - \pi(W, A; \hat{\alpha})\}^{-1}\{A - B(\hat{\eta})\} + B(\hat{\eta})]$. $[\hat{P}(A = 1 \mid C = 1)$ is a doubly robust estimator of $P(A = 1 \mid C = 1)$. That is, with $q(A, W)$ known, the estimator is consistent and asymptotically normal if either model $h(W; \alpha)$ or model $B(\eta)$ is correct. Final substantive conclusions depend on the set of functions $q(A, W)$ considered scientifically plausible (Robins, 2002). Robins, Rotnitzky, and Scharfstein (1999) showed this sensitivity analysis can be used as input for a full Bayesian analysis.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics* **113,** 231–263.

Baker, S. (2000). Analyzing a randomized cancer prevention trial with a missing binary outcome, an auxiliary variable, and all-or-none compliance. *Journal of the American Statistical Association* **95,** 43–50.

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

Gilbert, P., Bosch, R., and Hudgens, M. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59,** 531–541.

Hudgens, M. and Halloran, B. (2006). Causal vaccine effects on binary postinfection outcomes. *Journal of the Royal Statistical Society* **101,** 51–64.

Lerner, E. B., Billittier, A. J., Dorn, J. M., and Wu, Y. W. B. (2003). Is total out-of-hospital time a significant predictor of trauma patient mortality? *Academic Emergency Medicine* **10,** 949–954.

Lewis, D. (1973). Causation. *Journal of Philosophy* **70**, 556–567.

Little, R. and Rubin, D. (1999). Discussion of adjusting for non-ignorable drop-out using semiparametric nonresponse models by Scharsftein, Rotnitzky and Robin. *Journal of the American Statistical Association* **94,** 1121–1146.

Matsuyama, Y. and Morita, S. (2006). Estimation of the average causal effect among subgroups defined by post-treatment variables. *Clinical Trials* **2,** 1–9.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420. San Francisco, California: Morgan Kaufmann.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology* **17,** 276–284.

Quine, W. V. (1950). *Methods of Logic*. New York: Holt, Reinhardt and Winston.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Mathematical Modelling* **7,** 1393–1512.

Robins, J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease* **2**(suppl. 40), 139s–161s.

Robins, J. M. (1995). An analytic method for randomized trials with informative censoring: Part I. *Lifetime Data Analysis* **1,** 241–254.

Robins, J. M. (2002). Comment on "Covariance adjustment in randomized experiments and observational studies" by Paul R. Rosenbaum. *Statistical Science* **17,** 286–327.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3,** 143–155.

Robins, J. M. and Greenland, S. (2000). Causal inference without counterfactuals—Comment. *Journal of the American Statistical Association* **95,** 431–435.

Robins, J. M. and Rotnitzky, A. (2007). *On direct effects, surrogate markers and principal stratification.* Unpublished Technical Report. Department of Biostatistics, Harvard School of Public Health, Massachusetts.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, M. E. Halloran and D. Berry (eds), 1–92, IMA Volume 116. New York: Springer-Verlag.

Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine* **17,** 371–385.

Rubin, D. B. (2000). Causal inference without counterfactuals—Comment. *Journal of the American Statistical Association* **95,** 435–438.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31,** 161–170.

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

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Rejoinder to adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94,** 1121–1146.

Shepherd, B., Gilbert, P., Jemiai, Y., and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62,** 332–342.

Tan, Z. Q. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association* **101,** 1607–1618.

Tom Ten Have

Department of Biostatistics and Epidemiology
University of Pennsylvania School of Medicine
Blockley Hall, 6th Floor
423 Guardian Drive
Philadelphia, Pennsylvania 19104-6021, U.S.A.
email: *ttenhave@cceb.upenn.edu*

Frangakis et al. have presented a creative twist to the principal stratification approach in the context of a missing input due to death. For the first part of the paper extending to Section 5, the context is unique with the assumption that the intervention (Z) is 100% effective in preventing death (fully "protectable"). However, in Section 6, the authors go a long way to making the methodology more generalizable to contexts where the intervention is partially protectable, but which result in more identifiability problems. We now pursue with additional questions the authors' insightful comparison with other contexts, specifically the noncompliance randomized trial context after equation (2) and in Section 6. For both the fully protectable and partially protectable cases, we relate the authors' strategy to the noncompliance randomized trial context to better understand the ramifications of the assumptions. Following Frangakis et al., we make our comparisons of the assumptions in terms of the implications for the principal strata. The authors note that the four principal strata in their context (protectable always survivors, never survivors, and defiers) correspond in a one-to-one way to the four principal strata in the noncompliance, randomized trial context (compliers, always takers, never takers, and defiers).

Accordingly, the authors' interpretation of Assumptions 1, 2, and 3 in terms of principal strata can be compared to similar interpretations of analogous assumptions in the randomized trial context and corresponding principal strata. These randomized trial assumptions entail an ignorability assumption related to Assumption 1, the exclusion restriction, and a monotonicity assumption analogous to Assumptions 2 or 3 depending on the randomized trial design. The authors mention the relationship between the two contexts with respect to types of monotonicity assumptions. We now attempt to elaborate further on these relationships between the two contexts in terms of ignorability, exclusion restriction, and monotonicity assumptions.

1. Ignorability

The authors' Assumption 1 [A, $P \perp Z \mid X$] may be stronger than the ignorability or randomization assumption in the noncompliance, randomized context, where A occurs after Z and is measured for both levels of S. Randomization implies $A(1)$, $A(0)$, $P \perp Z \mid X$, where $A(1)$ and $A(0)$ are what A would potentially be if a subject were assigned to $Z = 1$ or $Z = 0$, respectively (e.g., Angrist, Imbens, Rubin, 1996). In the parlance of randomized trials with noncompliance, the ignorability assumption (Assumption 1) of the authors appears to say there is no overall intention-to-treat (ITT) effect of the baseline randomization (Z) on outcome (A). However, the ignorability assumption for randomized trials [$A(1)$, $A(0)$, $P \perp Z \mid X$] does not imply such a null effect of Z on A. We note that neither ignorability assumption implies a null ITT effect of Z on A within principal strata, which has implications for the ensuing discussion of the exclusion restriction.

2. Exclusion Restriction

The authors do not assume the exclusion restriction. On the face of it, one may ask if Assumption 1 implies the exclusion restriction, as defined in Angrist et al. (1996) for the noncompliance randomized trials context. However, under this exclusion restriction assumption, the ITT effects of Z on A equal zero in always and never takers (or never and always survivors), which is not necessarily true under Assumption 1. The ignorability assumption of Angrist et al. (1996) and exclusion restriction with a monotonicity assumption similar to Assumption 2 is sufficient for identifying the causal effect of treatment in compliers. One may ask if the exclusion restriction would make sense in the authors' context of missing input due to death, where A temporally precedes Z and S.

3. Monotonicity

We now attempt to elaborate on the authors' relation between Assumption 2 and monotonicity. Assumption 2 seems to lead to the converse situation of the Zelen single consent design (Zelen, 1990), under which controls do not have access to the randomized treatment, i.e., $\Pr(S = 1 \mid Z = 0) = 0$. In such cases, the always takers and defiers do not exist in comparison to the assumed nonexistence of the never survivors and defiers under Assumption 2 in the authors' context. That is, the protectable and always survivor principal strata are specified in the authors' case in contrast to the compliers and never-taker principal strata in the Zelen single consent design. Accordingly, the difference between the causal approach of the authors and the causal methods for the single consent design only involves differences between Assumption 1 versus the randomization assumption and the exclusion restriction assumption.

The authors emphasize the importance of some type of monotonicity assumption to identify causal effects in their context of missing input due to death. In both the full and partial preventability cases, they assume that defiers do not exist, as is often done in the noncompliance, randomized trials context. Under partial preventability when the never survivors exist and thus add parameters in need of identification, the authors impose parametric constraints involving covariates in equation (6) to identify the causal effects of

interest. The authors note however at the end of Section 5, "The fact that these quantities would be identifiable by our method even without the models in (6) if samples were large enough means that the results should not be sensitive to the particular parametric models, as long as they are flexible." Given this statement, how important are baseline covariates in identifying parameters under a fully parametric approach with partial preventability? It is clear that parametric relationships between P and baseline covariates X are very crucial for identifiability along with parametric distribution assumptions when monotonicity is relaxed (e.g., Rubin, 2004; Ten Have et al., 2004).

In the presence of protectables and always and never infected in the vaccine context, Gilbert, Bosch, and Hudgens (2003) augment the monotonicity assumption of no defiers with an additional but unidentifiable parametric relationship. Specifically, under the ignorability assumption $[A(0), A(1), P \perp Z \mid X]$, Gilbert et al. (2003) specify a parametric model relating $S(0)$ to $S(1)$ and A with the logistic function. In the Gilbert case, the log-odds ratio parameter corresponding to A is not identifiable. If Assumption 1 were to make sense in the vaccine context, would Assumptions 1 and 3 help preclude the need for such parameterizing such a relationship? Assumption 1 may not be feasible for the vaccine case, as it would imply that assignment to vaccine has no effect on disease level.

Finally, there are several interesting extensions of the authors' approach in their missing input/death context involving the incorporation of more information. Such information includes time to death (time to $S = 1$ since the intervening factor (Z)) and also multiple measurements of A across time some of which may be observed before $S = 1$. Given the popularity of joint survival/longitudinal outcome approaches and selection models, such extensions of the authors' work may be beneficial in the missing input context.

In summary, the authors' new implementation of the principal stratification approach has generated many interesting questions relating to other contexts and also challenges for incorporating additional information that may be helpful in identifying causal relationships between unmeasured and measured variables.

References

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

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in AIDS vaccine trials. *Biometrics* **59,** 531–541.

Ten Have, T., Elliott, M. M. J, Zanutto, E., and Datto, C. (2004). Causal models for randomized physician encouragement trials in treating primary care depression. *Journal of the American Statistical Association* **99,** 8–16.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31,** 161–170.

Zelen, M. (1990). Randomized consent designs for clinical trials: An update. *Statistics in Medicine* **9,** 645–656.

Yu Xie and Susan Murphy

University of Michigan
Quantitative Methodology Program, Survey Research Center
Institute for Social Research
426 Thompson Street
Ann Arbor, Michigan 48106-1248, U.S.A.

In "Principal Stratification Designs to Estimate Input Data Missing Due to Death," Frangakis, Rubin, An, and MacKenzie (hereafter FRAM) propose an analysis to do what may seem impossible: to recover input data that are missing due to death and then use the (observed and missing) input data to predict death. FRAM show that, under certain assumptions, this can be done with the introduction of an additional variable, "treatment," that possesses certain desirable properties.

We organize our comments as follows. First, we present the logic behind FRAM's analysis from the perspective of contingency table analysis. Second, with insights from this perspective, we will consider the implications of FRAM's analysis. Third, we discuss some considerations that should be taken into account in practice.

1. From Principal Stratification to Statistical Leverage

It appears that FRAM's analysis hinges on the notion of principal stratification (Angrist, Imbens, and Rubin, 1996; Frangakis and Rubin, 2002), i.e., the idea that discrete subpopulations, or strata, have distinct patterns of response to a treatment (called Z in the article). For simplicity, we focus on the main case discussed by FRAM: there are only two strata: a stratum of "always survivors" regardless of the treatment, and another stratum of "protectable" patients whose lives can be saved, but who cannot be harmed, by the treatment. Here the principal stratification assumption can be replaced by a less restrictive assumption:

Equation 2a: *If treatment is* $Z = 1$, *then the person must be alive at* 3 *months* ($S = 1$) *or, equivalently,* $P[S = 1 \mid Z = 1] = 1$.

Equation 2a is true if FRAM's assumption 2 is true, but Equation 2a invokes neither potential outcomes nor principal stratification. The crucial ignorability Assumption 1 of FRAM is that the assignment of Z is independent of both stratum membership and input data (A),

Table 1
Partial three-way crossclassified frequency table by Z, S, and A

Z	S	$A=0$	$A=1$	Subtotal
0	0	F_{000}	F_{001}	F_{00+}
	1	F_{010}	F_{011}	F_{01+}
1	0			
	1	F_{110}	F_{111}	F_{11+}

Note: Entry F_{ijk} refers to the frequency crossclassified by $Z = i$, $S = j$, and $A = k$. Cells shaded light gray are not observed but estimated. Cells shaded dark gray are not allowed by assumption.

conditional on covariate X (see below for more on this assumption).

We note that covariate X plays no special role in FRAM's article except to make the ignorability assumption plausible. Thus, the discussion that follows is conditional on X. In terms of time ordering, the input data, A, exist prior to the critical event (here an injury), the treatment Z occurs shortly after the injury but prior to death, and S denotes death (here coded as 1 if the subject is alive 3 months after the injury and 0 otherwise). Note that S is always observed so S is S^{obs} in the FRAM article, and furthermore $S = ZS(1) + (1 - Z)S(0)$, where $S(z)$ denotes the potential outcome when the treatment $Z = z$. Z is always observed, but A is observed only if $S = 1$.

Because Z, S, and A are all binary, we can capture their joint distribution with a three-way crossclassified contingency table, shown in Table 1. We use F_{ijk} to denote the frequency count in the crossclassified table for the cell $Z = i$, $S = j$, and $A = k$, with $i = 0, 1$, $j = 0, 1$, and $k = 0, 1$. We use the plus sign, "+," in the subscript to denote the subtotal for summation over a particular subscript. Two features stand out in Table 1. First, because all patients who received the treatment ($Z = 1$) survived, the third row (representing $Z = 1$, $S = 0$) contains structural zeros. Second, while we know the subtotal of the first row, representing the situation of $Z = 0$, $S = 0$, we do not know the distribution of A in that row. Indeed, recovering this distribution from patients who had died before the interview is a main research objective here. Due to these two unique features, Table 1 differs from the usual $2 \times 2 \times 2$ contingency table. We call such a table as Table 1 a "partial contingency table."

How can we recover the distribution of A in the row in the partial contingency table? We make use of the ignorability assumption in FRAM's approach and our Equation 2a. Equation 2a sets the third row ($Z = 1$, $S = 0$) to structural zeros so that $F_{1+0} = F_{110}$, and $F_{1+1} = F_{111}$. The independence assumption for the relationship between Z and A means that the odds of $A = 1$ versus $A = 0$ is the same across the two different values of Z. We thus have the following constraint:

$$F_{1+1}/F_{1+0} = F_{111}/F_{110} = (F_{001} + F_{011})/(F_{000} + F_{010}). \quad (1)$$

We then add to equation (1) the known information that

$$F_{000} + F_{001} = F_{00+}. \quad (2)$$

Table 2
Numerical example using the NSCOT data for the partial three-way crossclassified frequency by Z, S, and A, by covariate X

Z	S	$A=0$	$A=1$	Subtotal
X = low-injury severity				
0	0	3	14	17
	1	257	72	329
1	0			
	1	6	2	8
X = high-injury severity				
0	0	18	6	24
	1	95	5	100
1	0			
	1	10	1	11

Notes: Cells shaded light gray are not observed but estimated. Cells shaded dark gray are not allowed by assumption. Estimated relative risk of death is 14.1.

Cells shaded light gray are not observed but estimated. Cells shaded dark gray are not allowed by assumption. Estimated relative risk of death is 3.4.

We can easily solve equations (1) and (2) for two unknowns, F_{000}, F_{001}. In Table 2, we present our numerical results based on the information provided by FRAM for their data from the National Study on the Costs and Outcome of Trauma Centers (NSCOT). There may be small discrepancies between our results and the actual results, because we recovered counts from FRAM's original results in percentages. Following FRAM, we also treat the illustrative example as if we have population data and thus do not consider statistical inference issues.

From the approach of a contingency table analysis, we see why FRAM's analysis works. We think that our contingency table approach is more intuitive and more straightforward. One advantage of our approach is that equation (1) clearly reveals how the missing information pertaining to the distribution of A for the dead group ($Z = 0$, $S = 0$) is recovered: it compensates the distribution of A among untreated survivors ($Z = 0$, $S = 1$) so that the combined distribution equals that of the treated group ($Z = 1$). Everything else being equal, the distribution of A among the dead ($Z = 0$, $S = 0$) moves in the same direction as the distribution of A in the treated group ($Z = 1$) and in the opposite direction from that of the distribution of A among untreated survivors ($Z = 0$, $S = 1$). We are clearly borrowing information from other related groups. It is as though we are able to move an enormous object by a mechanical lever. Thus, FRAM's approach is an exemplary case of using "statistical leverage."

2. Implications for Research Objectives

In FRAM's analysis using statistical leverage, an additional treatment variable can recover the missing information about input data. We showed earlier that we were able to fill in the cells of missing data in Table 2 for their numerical example. How well does the recovered information serve the original objectives of the substantive research? To answer this question, let us visit the research objectives that FRAM's analysis is

intended to help achieve. The abstract clearly states the two research objectives: (i) "to measure 'input' variables, which describe the period before the critical event, and to characterize the distribution of input variables in the cohort"; and (ii) "to measure 'output' variables, primarily mortality, after the critical event, and to characterize the predictive (conditional) distribution of mortality given the input variables in the cohort."

If we are to take the first objective literally, it is not necessary to fill in the missing data, as we did in Table 2. By assumption, the distribution of the input variable (A) is independent of Z. Thus, the distribution of the input variable (A) conditional on Z also describes the unconditional distribution of the input variable (A), as the following is true by the ignorability assumption (Assumption 1):

$$P(A = 1 \mid Z = 1) = P(A = 1 \mid Z = 0) = P(A = 1). \quad (3)$$

Of course, this does not tell us $P(A \mid Z = 0,\ S = 0)$, which can only be recovered after missing values are estimated.

Achieving the second research objective requires an additional assumption; here we use Equation 2a. If we take the stated objective literally, the researcher is interested in the following quantities for the entire population:

$$P(S = 0 \mid A = k), \quad k = 0, 1. \quad (4)$$

We can further decompose these quantities by treatment status (Z):

$$\begin{aligned} &P(S = 0 \mid A = k) \\ &\quad = P(S = 0 \mid A = k, Z = 0)P(Z = 0 \mid A = k) \\ &\qquad + P(S = 0 \mid A = k, Z = 1)P(Z = 1 \mid A = k), \\ &\quad = P(S = 0 \mid A = k, Z = 0)P(Z = 0) \\ &\qquad + P(S = 0 \mid A = k, Z = 1)P(Z = 1) \\ &\quad = P(S = 0 \mid A = k, Z = 0)P(Z = 0). \end{aligned} \quad (5)$$

Note that we obtained the second line of equation (5) by using the independence assumption and the last line of equation (5) by using the information that all subjects survive if treated ($Z = 1$). Because $P(Z = 0)$ is unrelated to A, this term is cancelled in the formula for the relative risk, the ratio of conditional probabilities:

$$\begin{aligned} &[P(S = 0 \mid A = 1)]/[P(S = 0 \mid A = 0)] \\ &\quad = [P(S = 0 \mid A = 1, Z = 0)]/[P(S = 0 \mid A = 0, Z = 0)]. \end{aligned} \quad (6)$$

Equation (6) can be estimated using our partial contingency table approach by

$$[F_{001}/(F_{001} + F_{011})]/[F_{000}/(F_{000} + F_{010})]. \quad (7)$$

We present our numerical results using equation (7) for the illustrative example.

Two comments concerning the second research objective are in order. First, if we wish to know the mortality rates by the values of the input variable, it is necessary to know the proportion not receiving treatment in the population, $P(Z = 0)$. When the researcher is interested only in the relative risk, or odds-ratio, by the input variable, $P(Z = 0)$ can be ignored. Second, the appearance that the group of treated persons ($Z = 1$) do not seem to affect the relative risk in equation (6) is misleading, as these persons affect the estimation of the missing information as part of the "statistical leverage" discussed earlier.

3. Practical Considerations

Although FRAM's analysis allows the researcher to uncover missing data that are not missing at random through the power of statistical leverage, implementation is not trivial. Below, we discuss some considerations that researchers should take into account when adapting the analysis in practice.

First of all, the researcher needs to carefully consider the treatment variable Z. A number of questions arise:

1. Is Z an existing treatment in practice or a new intervention as part of the study design?
2. If the researcher does not manipulate Z, are we comfortable with the assumption that Z and A are independent conditional on covariates?
3. If the administrator knows the effectiveness of Z, what prevents her/him from "overprescribing" the treatment to reduce deaths?
4. Does the effectiveness of Z vary with time, location, population, or the proportion being treated?

While the first two questions are straightforward, as they are concerned with the ignorability assumption, the last two questions need some discussion.

Let us generalize the idea of principal stratification. Suppose the population is not divided into two strata—those who always survive and those who are helped by treatment—but numerous subclasses characterized by the degree to which treatment Z helps survival. That is, the counter-factual response function for person i is a continuous score, depending on the person's latent response function R, $R_i = S_i(1) - S_i(0)$. Under the common assumption of monotonicity (Angrist et al., 1996; Frangakis and Rubin, 2002), we specify that $0 \leq R_i \leq 1$. Further imagine that because the administrator of Z knows additional information (unknown to the researcher) about patients' and hospitals' conditions, he or she would assign Z to those patients who would benefit most from the treatment. That is, we entertain the possibility that the likelihood of receiving Z is correlated with the amount of treatment effect R. When this is the case, increasing the proportion of Z necessarily results in lowering the average treatment effectiveness of treatment Z, as the composition of the stratum receiving treatment ($Z = 1$) has changed from having a higher average R score towards having a lower average R score (i.e., from benefiting more on average to benefiting less on average). This discussion illustrates a practical difficulty with the principal stratification approach in general: we do not know individuals' memberships in the various strata, as the existence of the strata can only be inferred from the group level. Thus, we may view principal strata either

as distinct subpopulations with distinct response patterns or as aggregations of heterogeneous individuals with somewhat similar response patterns. The latter, nominal perspective is consistent with the view of heterogeneous treatment effects at the individual level. Our concern is that if we accept the nominal perspective, policy or technological changes can change the proportion and at the same time the composition of the group of subjects receiving treatment. Properties of principal strata, nominally defined, are thus not fixed and are subject to change.

We next consider the role of the covariates X. From the perspective of assumptions needed to make FRAM's analysis work, X precedes both A and Z and indeed makes them independent of each other conditional on X. From the perspective of data collection, X was not provided in the interview, as it would, like A, then be truncated by death. Conceptually at least, one would like to condition on a rich set of covariates before accepting the conditional independence assumption. For example, we would like to know a person's medical history, demographics, and family socioeconomic status. Needless to say, it is not possible to condition on them if they are considered part of A instead of X. In other words, an input variable A and a covariate X differ in two respects: (i) X is observed, whereas A is only partially observed; (ii) X is to be conditioned on, whereas A and Z are assumed to be conditionally independent. Strict association of partial observability with the conditional assumption is more a practical convenience than a necessary condition justified by science. Conceptually at least, it is possible that we may wish to condition on covariates that may only be partially observed. However, not observing them in practice would force us to convert them into input data (A) that would then need to satisfy the independent assumption.

There is no easy and magic solution to this problem. We recommend that the researcher collect more and better data as a possible remedy. One possibility is to use administrative records (such as the death certificates and medical records). Another possibility is to interview surviving family members for proxy reports. In general, better data can yield far more statistical information than can be achieved through statistical leverage. In the approach of pushing for better-observed data, the boundary between input data and covariates is blurred.

4. Conclusion

The FRAM analysis is intuitively appealing, and relatively easy to implement. One of the most interesting features of the analysis is that it allows the researchers to impute data that do not satisfy the ignorability assumption alone, but under a *model* that satisfies ignorability.

If the input data were to satisfy the ignorability assumption, the distribution of the input data would be the same between survivors and nonsurvivors. This is clearly implausible and is rejected by FRAM. Even after introducing a new treatment, FRAM do *not* assume ignorability in the distribution of the input data between survivors and the nonsurvivors within treatment status. Rather, the ignorability assumption is imposed on the two-way marginal association between the treatment variable and the input variable. This restriction allows FRAM to recover missing input data among nonsurvivors.

How well FRAM's analysis will work in practice is a substantive question that will depend on concrete applications. At the minimum, the new analysis provides alternative estimates so as to characterize the distribution of input data and the association between the input data and the risk of deaths. This exercise is informative even if one does not necessarily believe that the underlying model is correct, for the alternative estimates provide some sensible and plausible bases for the researcher to critique and improve upon. For this and many other reasons previously discussed, we recommend this article to all who are interested in the topics it covers: missing data, causal inference, principal stratification, and partial contingency table.

References

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

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

Rejoinder

Constantine E. Frangakis,
Donald B. Rubin,
Ming-Wen An,
and Ellen MacKenzie

We thank the editors for the opportunity to have the topic of this article discussed. We also thank the discussants for their generally interesting comments.

1. Xie and Murphy

Xie and Murphy (XM) describe our problem using only observed-data representations, and then discuss some additional practical issues.

Physics versus pure empiricism. XM present a reduced form of our problem using only the resultant observed data. Motivated from this representation, XM (and also Robins, Rotnitzky, and Vansteelandt—RRV) suggest that one does not need to invoke potential outcomes and principal strata. We disagree: Potential outcomes and principal strata are essential in order to *formulate* the problem and goal, to state explicit assumptions (such as ignorable treatment

assignment), and to *devise possible designs* to address the problem. For instance, with only notation for observed data, it is not possible even to define the meaning of a causal effect. That meaning was central in our approach—to regard a value as truly missing (i.e., not observed but observable), only if there exists, in principle, an intervention that would have caused it to be observed. The reader can appreciate the need for potential outcomes also from XM's own writing: when they comment outside of our specific problem, they too invoke potential outcomes and principal strata (see their discussion after question (d) in "practical considerations," where their R_i is defined as the difference of never jointly observable potential outcomes: $S_i(1) - S_i(0)$).

More generally, a representation in terms of potential outcomes and principal strata is required if one is to describe the theoretical, physical underlying system of the problem. Many analogies regarding such physical versus purely empirical representations can be drawn. For example, man went to the moon based on Newton's theoretical, *physical* (albeit not quite correct) model of nature's laws. That voyage would not have been possible if Newton had not persisted in seeking a physical model, but instead had proposed—and if we had accepted as appropriate—some nondifferentiable step function (e.g., based on a CART—tree diagram) that would stop after "explaining" empirically only his discrete, few observations.

In summary, postulating a theory in terms of its underlying physics has been, and will continue to be more beneficial than mere explanation in terms of observed data, because a physical system is actually more parsimonious and thus more generalizable, and hence more powerful for predicting other observable events.

On practical considerations. A researcher needs to consider the thoughtful questions (a)–(d) that XM raise, and address them based on the ability to obtain data on factors approximately satisfying our assumptions. An example is question (c): if the prevention factor z is known to be effective, why does the decision maker not administer the most effective level of z to all? The answer involves obstacles external to the decision maker. Taking, for example, the time to transport an injured patient to the hospital, and adjusting for severity of injury and knowledge that time is important, considerable variation in time can still exist because of other factors: how promptly the injury victim was first spotted and reported; how close the nearest help was; availability of fast transport at the time; and traffic and other problems encountered by the transport mode. This comment also addresses RRV's point on ethical considerations: variation in such obstacle factors cannot be generally viewed as ethical or not, because these obstacles are rarely in the control of the ethically charged decision maker for z. Of course, Z_i is assigned by the decision maker so as to maximize the *anticipated* likelihood of survival, but this likelihood is only conditional on what the decision maker knows, and so after we condition on that knowledge, we can effectively assume ignorability.

Regarding XM's discussion of more general principal strata, certainly the meaning of the strata $S_i(z)$ depend on the meaning of the prevention factor z, but this is not a complication of principal strata, but a consequence of meanings changing with problems. Within a problem, though, the meaning of $S_i(z)$ *does not* depend on the assignment mechanism for the actual levels Z_i.

XM wonder about the distinction between the covariates we denoted as X and the input factor A, stating that "X precedes both A and Z." This is not generally correct. Some covariate values are determined prior to both A and Z, such as age or gender, but other covariate values, and often those used to ensure ignorability, are determined prior to Z but after A. In our example, X was the severity of injury as judged by the medical personnel *after* the injury occurred, whereas the input variable A was a disability whose value is determined *before the injury*, but only recorded at the interview after the injury.

More important, as we have emphasized in the article, there is a clear *scientific* distinction between the critical covariates X and the input A: the covariates X used to ensure ignorability need only be those that were involved in the decision maker's informed choice to administer or not the prevention factor (for example, X can often leave out factors causing variation in z such as the obstacle factors just described). The key fact that makes it *easier to record* X than A is this: If the decision maker for z is a person *other* than the injured victim, we can, in principle, talk to that decision maker (whether or not the victim eventually dies) and ask for the value of *all* those variables X that the decision maker used for the assignment of the prevention factor. We cannot do the same for A because, by definition, its accurate measurement depends on the victim's ability to be interviewed, which is impossible if the victim dies.

2. Ten Have

Ten Have commented on the role of an exclusion restriction and the role of covariates, and has indicated numerous directions for possible fruitful extensions to our methods.

On ignorability and exclusion. Ten Have wonders about our assumption of ignorability, that is, $(A, P) \perp\!\!\!\perp Z \mid X$, and its relation to exclusion restrictions typically made in settings of noncompliance. Because the factor A is, by design, an input factor that precedes the prevention factor z, the value of A cannot be changed (for any person) by changing the level of z. If we had allowed potential outcomes for A under $z = 0, 1$, i.e., $A_i(0)$, $A_i(1)$, then the exclusion restriction $A_i(0) = A_i(1)$ would have been a *consequence* of the temporal ordering of the design, and not an assumption, and that is why we need not make it.

Now, given that A precedes z, it follows that A, like any other covariate, will be balanced between levels of z after we condition on the variables that were used to make the assignment of the actual levels Z_i—hence the ignorability assumption 1. For that assumption, we disagree with Ten Have's claim that "neither ignorability assumption implies a null ITT effect of Z on A within principal strata": it is not difficult to show that the ignorability assumption 1, i.e., $(A, P) \perp\!\!\!\perp Z \mid X$, implies that $A \perp\!\!\!\perp Z \mid P, X$.

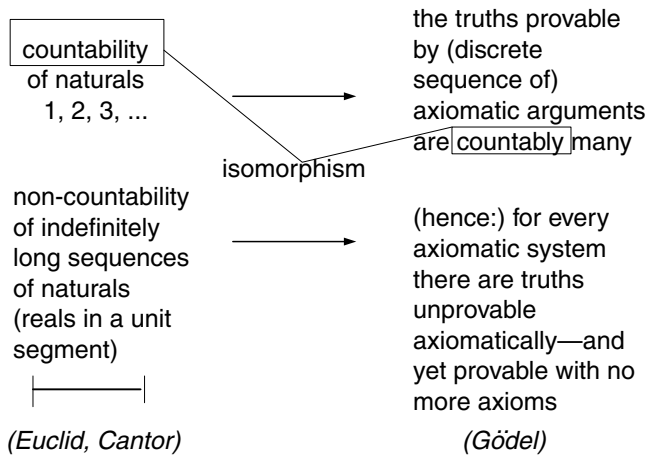

Figure 1. A remarkable case of isomorphism.

On relations to other problems—what is important in an isomorphism? Ten Have observes the similarity of the mathematics in our problem to the mathematics of noncompliance and Zelen's design, a comment related to the isomorphism discussion of RRV.

It is revealing to explicate the source of importance in an isomorphism. To do so, we will invoke a striking example—the isomorphism involved in Gödel's theorem of incompleteness (Gödel, 1931; see also Nagel and Newman, 2001). In brief, Gödel considers axiomatic systems, where axiomatic proofs are constructed by logically building those proofs based on the axioms. Assuming the system is consistent, Gödel then constructs a proposition that has a remarkable duality: (1) there exists no axiomatic proof that the proposition is true, yet (2) there exists a nonaxiomatic, but fully valid, proof that the proposition is true, without invoking additional axioms. The relation to our discussion is this: Gödel's theorem was a completely *new* (and unexpected) result, yet it was isomorphic to another well-known result, that indefinitely long sequences of natural numbers are uncountable. A somewhat liberal, but useful, explanation of the mapping, provided in the figure, is that the proofs that we can construct axiomatically using finite sentences must be countably many, yet there are uncountably many truths (Figure 1).

Thus, the critical element in any isomorphism is the intuition that leads us to *see, in the first place,* how one problem—in our case, missing data—could be solved by a design that draws power from an *appropriate* isomorphism—in our case, involving causal inference. The isomorphism here is not really about why instrumental variables work, but much deeper; it involves the relation, discussed in the final section of the target article, between the situation where a quantity can be legitimately viewed as missing, and the requirement that there should exist, at least in principle, an intervention that could have made that quantity observed.

On monotonicity and covariates. Ten Have also asks how important covariates are for identifying parameters under a fully parametric approach. Under the assumptions that allow such identifiability, covariates are as important in the parametric as in the nonparametric formulation, because in order to justify the ignorability assumption 1, we must condition on the covariates involved in the decision maker's choice to administer the prevention factor z. Of course, if monotonicity is relaxed, covariates that predict the *direction of effect of the prevention factor* are also important for narrowing the ranges of the plausible distributions, as Ten Have suggests.

Moreover, Ten Have points toward the connections to the very interesting problem of evaluating a vaccine's efficacy on the viral load for post-randomization infectees—a problem in which the use of principal stratification was initiated by Gilbert, Bosch, and Hudgens (2003). The scientific structure of that problem, as Ten Have observes, is different from this one because an exclusion restriction in the vaccine problem is questionable and should not be assumed a priori.

Finally, it is rewarding to see how quickly Ten Have points to many new fruitful directions and challenges in which such ideas can be useful.

3. Robins, Rotnitzky, and Vansteelandt

RRV mainly comment on our assumptions and goal. Their comments about our assumptions are addressable. Their comments about our goal are not relevant to a researcher who interprets the meaning of our result in a scientific context.

(i) Addressing RRV's points on assumptions. RRV's comments about our assumptions are addressable because they are assessable. For example, RRV say that physicians know that death can arise from hemorrhage when a thrombolytic drug is given [*versus if it is not given*] after an infarction. That is true but is not relevant to our assumption regarding *timing* of administration: Physicians also know that among infarction victims to whom a thrombolytic drug *is administered*, the sooner the drug is given, the higher the likelihood of survival; in fact there is strong evidence from randomized trials that the probability of hemorrhage is practically zero (est. at 0.2%) when the drug is given within 2 hours after a myocardial infarction, whereas when the drug is given later than 2 hours, the probability of hemorrhage is estimated at 2.5%, a relative risk of more than 10 fold (Steg et al., 2003). Thus the note of RRV is incomplete and could mislead the casual reader.

RRV also point to the debate about whether attempting to stabilize injured persons is better before or after transporting them to the hospital. This debate exists but is also not relevant to our assumptions. Obviously those who transport injured victims know whether or not they made such stabilizing attempts. Therefore, when this information is used (e.g., as a stratifier in the variables X that were used to make the decisions regarding transportation), its variation is controlled (in principle) and is no longer a concern. With analogous thought

and adjustments, one can address RRVs other concerns.

(ii) The importance of using scientific meaning in our goals. It is important to recall the two observations we made under partial preventability (Section 5.1). First, we stated that we cannot fully estimate the predictive distribution of mortality given the input factor. Second, we showed that we can still estimate the distributions of the input variable, A, for patients who are protectable by the prevention factor, which also implies we can estimate the distribution of A for patients who are "always survivors." That is, we can estimate

$$P(A_i = 1 \mid P_i = \text{ protectable}) \text{ and}$$

$$P(A_i = 1 \mid P_i = \text{ always survivor}). \quad (1)$$

RRV's comments on this point (regarding the anti-bird-flu drug) essentially repeat—but stop at—our first observation, without paying attention to our second observation. The strata "protectable" and "always survivors" have scientific *meanings*, and are not merely technical abstractions to be averaged over (or not) depending on some statistical goal. In most problems, the strata of "protectable" and "always survivors" are best understood as gradations of a condition in a single underlying system. For example, in the injury problem, this system determines how much injury damage a person can endure at various levels of a treatment factor: An "always survivor" is a person of higher endurance than a "protectable" person, who needs the more effective level of the treatment to survive. In RRV's example, the system determining "protectable" and "always survivor" is the immune system, and "always survivors" are those with generally stronger immune system than the "protectables."

Using the scientific meaning of the principal strata is crucial because it allows us to interpret the estimable comparison in (1) above: if "always survivors" have a higher proportion of the input factor's state $A = 1$ than the "protectables," then this implies that state $A = 1$ is associated with more robust states of the system that determines the principal strata. For example, it would imply that $A = 1$ is associated with higher endurance to injuries, in the injury example; and that $A = 1$ is associated with higher immunity to the virus, in the bird flu example. This conclusion is reachable without needing to identify fully the predictive distribution of mortality, although there is the need to think about the meaning of the principal strata.

Of course, the association between the input factor and the principal strata of protectable versus always survivor does not imply causality. RRV discuss this at length (in their discussion of the psychiatrist's hypothesis), but this issue seems to be entirely obvious. Our goal—to learn about the associations of A and mortality when A is missing under death—provides information to suggest that A is related to (and thus *may be* causally involved in) the system determining the endurance of a person to survive an injury. With this suggestion established, whether the factor A is or is not a causal agent must be addressed with a different design, and cannot be addressed simply by the notational arguments of RRV.

(iii) On statistical issues. RRV also reiterate at length our observation that our problem is related, indeed isomorphic, to other problems involving principal stratification. Although we also could have detailed many additional examples from our own work (e.g., Frangakis et al., 2004; Li and Frangakis, 2006; Rubin, 2006; Jin and Rubin, 2007), we believe that is more beneficial to the reader to read our conceptual rejoinder to Ten Have on isomorphisms.

(iv) The role of objectivity and sensitivity analysis. We agree that sensitivity analyses can be useful, but the question is how to conduct them. A sensitivity analysis can only be useful to the extent that the framework in which it is formulated is rich enough so that it can provide, at least partly, an objective assessment of the values or ranges of the sensitivity parameters. "Objective" here does not mean "absolutely correct," but it does means "based on assumptions that are understandable," because then, as we have seen from discussions like this (item (i)), one can clarify the ways to assess and address concerns about these assumptions. It is such objectivity that our design and methods provide.

Acknowledgements

We thank again the editors and discussants for the generally stimulating exchange of ideas, and Spyridon Kotsovilis for insightful discussions on Gödel's theorem.

References

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Mahboobeh, S., Vlahov, D., and Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association* **99,** 239–249.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in AIDS vaccine trials. *Biometrics* **59,** 531–541.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I. *Monatshefte für Mathematik und Physik* **38,** 173–198.

Jin, H. and Rubin, D. B. (2007). Principal stratification for causal inference with extended partial compliance: Application to Efron-Feldman Data. To appear in the *Journal of the American Statistical Association.*

Li, F. and Frangakis, C. E. (2006). Polydesigns in causal inference. *Biometrics* **62,** 343–351.

Nagel, E. and Newman, J. R. (2001). *Godel's Proof.* New York: New York University Press.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with censoring due to death (with discussion). *Statistical Science* **21,** 299–321.

Steg, P. G., Bonnefoy, E., Chabaud, S., Lapostolle, F., Dubien, P. Y., Cristofini, P., Leizorovicz, A., Touboul, P.; Comparison of Angioplasty and Prehospital Thrombolysis In acute Myocardial infarction (CAPTIM) Investigators. (2003). Impact of time to treatment on mortality after prehospital fibrinolysis or primary angioplasty: Data from the CAPTIM randomized clinical trial. *Circulation* **108,** 2851–2856.