 Vol 31: 161–170, 2004

Direct and Indirect Causal Effects via Potential Outcomes*

DONALD B. RUBIN

Harvard University

ABSTRACT. The use of the concept of 'direct' versus 'indirect' causal effects is common, not only in statistics but also in many areas of social and economic sciences. The related terms of 'biomarkers' and 'surrogates' are common in pharmacological and biomedical sciences. Sometimes this concept is represented by graphical displays of various kinds. The view here is that there is a great deal of imprecise discussion surrounding this topic and, moreover, that the most straightforward way to clarify the situation is by using potential outcomes to define causal effects. In particular, I suggest that the use of principal stratification is key to understanding the meaning of direct and indirect causal effects. A current study of anthrax vaccine will be used to illustrate ideas.

Key words: anthrax vaccine, biomarkers, causal inference, principal stratification, Rubin Causal Model, surrogate outcomes

1. Introduction – direct and indirect causal effects

Causal inference is an area of rapid and exciting development and redevelopment in statistics. Fortunately, the days of 'statistics can only tell us about association, and association is not causation' seem to be permanently over. The particular topic of this presentation concerns the concept of 'direct versus indirect causal effects'. The use of this concept is common in the economic and social sciences, e.g. the effect of wealth on health and health on wealth (for my views, see the discussion by Mealli & Rubin, 2003 of Adams *et al*., 2003). In the biomedical and pharmacological sciences, there are the closely related concepts of 'biomarkers' and 'surrogate outcomes', and the related language describing an intermediate outcome that is 'on the causal pathway' to the final outcome. A generally unappreciated role is that of covariates, pretreatment variables unaffected by treatment assignment, but I will touch on their importance only at the end, after describing fundamental issues.

Some current statistical discussion of direct and indirect causal effects is in terms of graphical models (e.g. explicitly Pearl, 2001, and the discussion by Lauritzen of this article). This use is especially common in longitudinal settings. Such graphical displays, with their nodes, directed arrows, undirected arrows, absence of arrows, etc. are quite seductive. Nevertheless, I still feel as I did in my discussion with Imbens (Imbens & Rubin, 1995) of Pearl (1995), that the framework is inherently less revealing than the potential outcomes framework because it tends to bury essential scientific and design issues. The potential outcomes framework is sometimes called the 'Rubin Causal Model' (RCM) (Holland, 1986), but it has roots in the context of randomized experiments with randomization-based inference in the work of Neyman (1923) and Fisher (1925). The term RCM comes from extensions to observational studies (e.g. Rubin, 1974, 1977) and other forms of inference (e.g. Bayesian – Rubin, 1978); see Rubin (1990) on Neyman (1923).

*This paper was presented as a Specially Invited Paper at the 19th Nordic Conference on Mathematical Statistics, Stockholm, June 2002 (NORDSTAT 2002).

Despite criticisms I may have of the graphical approach, current graphical approaches seem to be a clear advance with respect to causal inference over older, less subtle graphical approaches.

The general theme here is that the concepts of direct and indirect causal effects are generally ill-defined and often more deceptive than helpful to clear statistical thinking in real, as opposed to artificial, problems. A clear conceptual formulation of the relevant issues for practice must precede a correct mathematical statistical attack on these issues. I feel that a complex causal problem is typically understood using the potential outcomes framework (e.g. noncompliance in Angrist *et al.*, 1996) and, then, after the correct conceptual structure is seen, it is subsequently translated into a graphical language, if possible. These are controversial statements made partially to stimulate clarifying rebuttal.

This paper is not designed to be a systematic review of the literature on the topic of direct and indirect causal effects. It is a presentation of my current work and views on these issues that have evolved over more than three decades. Section 2 briefly reviews the RCM, and section 3 introduces the topic of direct and indirect causal effects using a very simplified version of anthrax vaccine experiments. Section 4 describes the plan for using immunogenicity measurements as a surrogate for survival after exposure to lethal doses of anthrax. Section 5 describes observed data in the experiment and how naive uses of the data on immunogenicity lead to incorrect conclusions. Thinking about how to multiply impute the missing potential outcomes, described in section 6, clarifies the situation, and section 7 expands this discussion by giving illustrative assumptions that can make imputations more precise.

2. Potential outcomes and the assignment mechanism – the RCM

The key organizing principle for addressing the topic of direct and indirect causal effects will be based on the concept of 'principal stratification' (Frangakis & Rubin, 2002). This perspective can be viewed as having its seeds in the 'instrumental variables' method of estimation, as described within the context of the RCM in Angrist *et al.* (1996) from the frequentist perspective and Imbens & Rubin (1997) from the Bayesian perspective, but having roots in work by economists such as Tinbergen (1930) and Haavelmo (1944).

One of the tenets of the RCM perspective is that only limited progress is possible without taking as fundamental the definition of causal effects as the comparison of potential outcomes on one common set of units (i.e. *not* the comparison of the treatment potential outcomes for one set of units and the control potential outcomes for a different set). A second tenet of this perspective is the need, when drawing causal inferences, to posit an assignment mechanism, a model for how units were assigned the treatments they received. A third tenet is that we need to be explicit about assumptions because human beings, naturally, are *very* bad at dealing with uncertainty, and repeatedly fall prey to 'paradoxes' such as 'Simpson's Paradox' (Simpson, 1951).

Display 1 presents data from a hypothetical comparison of a new surgery relative to a standard surgery using the potential outcome notation: X is a covariate (age) unaffected by the assigned treatment; $W = 0,1$ is the treatment assigned ($W_i = 1 =$ new, $W_i = 0 =$ standard); $Y(0)$ is the value (years lived post operation) that would have been observed under the control (old) operation; and $Y(1)$ is the value that would have been observed under the experimental (new) treatment. Throughout, we accept the stable unit-treatment value assumption (SUTVA) (Rubin, 1980, 1990), so that Display 1 represents all values that could have been observed under any assignment of old and new operations.

First notice that, on average, or 'typically', the old (standard) operation is better than the new operation: this is true for five of the eight patients, and true overall as well when assessed

Display 1. *Illustrative example of need for assignment mechanism*

Covariate X	W	Potential outcomes $Y(0)$	Potential outcomes $Y(1)$	Individual causal effects $Y(1) - Y(0)$
68	1	13	14*	+1
76	0	6*	0	−6
66	0	4*	1	−3
81	0	5*	2	−3
70	0	6*	3	−3
72	0	6*	1	−5
81	1	8	10*	+2
72	1	8	9*	+1
True averages		7	5	−2
Observed averages		5.4*	11*	

*Observed values. (Y, years lived postoperation; X, age at start of study).

by the mean or median individual causal effect, or the comparison of median potential outcomes under the two operations. Secondly, notice that, at least in this study, all patients received the optimal treatment for themselves. If this doctor always were able to do this (a perfect doctor), this is a doctor that we would all love to have for our own health care. In this case, the assignment mechanism would be:

$$P(W|X, Y(0), Y(1)) = \prod_{i=1}^{8} P(W_i = 1|X_i, Y_i(1), Y_i(0)),$$

$$\text{where } P(W_i = 1|X_i, Y_i(1), Y_i(0)) = \begin{cases} 1 & \text{if } Y_i(1) \geq Y_i(0) \\ 0 & \text{otherwise.} \end{cases}$$

Suppose this assignment mechanism is the one the doctor used.

Thirdly, notice that the observed data appear to lead to the opposite conclusion from the truth: the three patients treated with the new operation all live longer than the five patients treated with the standard operation and, moreover, there is no overlap in the distribution of the observed outcomes. That is, the least successful person treated with the new operation lives longer after the operation than the most successful person treated with the standard operation. The implicit conclusion about the efficacies of the two operations is wrong, and formal statistics can help by showing that the simple comparison of outcomes under new and standard operations is predicated on the underlying assignment mechanism being one that has had the units randomly assigned to treatment and control, which was not done. For a completely randomized experiment with three treated and five control units we have:

$$P(W|X, Y(0), Y(1)) = \begin{cases} 1/56 & \text{if } \sum W_i = 3 \\ 0 & \text{otherwise.} \end{cases}$$

For causal inference, we require a model for the assignment mechanism, that is, we need a model for $\Pr(W|X, Y(0), Y(1))$. The more typical models, and those implicit in most graphical approaches, are models for the 'data', $\Pr(X, Y(0), Y(1))$, which generally are not necessary, as documented by the voluminous literature on randomized-based analyses of randomized experiments, from Neyman (1923) to Frangakis & Rubin (1999), which develops causal inferences based solely on the assignment mechanism. Thus we need to posit a model for the assignment mechanism (e.g. the experimental design), but we are inherently drawn to a model for the data (i.e. the science) – hence the seductiveness of graphical models. The approach of

modelling the data and ignoring the assignment mechanism only works generally in the absence of uncertainty (e.g. the new operation is *always* 2 years better than the old operation and a randomized experiment was conducted).

To introduce the topic of indirect and direct effects, I now turn to a real example of some importance.

3. Randomized trials of anthrax vaccine

Two sets of randomized placebo-controlled trials on anthrax vaccine are being conducted by the United States Centers for Disease Control and Prevention (CDC). The trials have two purposes: one is to find the dosing regimen of vaccination that is effective for humans when exposed to lethal doses of anthrax, and the second purpose is to find the regimens that are safe in the sense that negative side-effects of the vaccine are minimal. I have been actively involved in the design and analysis of both trials.

The first trial has six arms for different vaccination regimens, including a placebo arm, and the second trial is parallel to the first. The difference between them is that the first involves human volunteers with extensive reactogenicity measurements (e.g. side-effects), as well as extensive immunogenicity measurements (e.g. blood antibody levels), whereas the second involves macaques (rhesus monkeys) with parallel immunogenicity measurements, plus the outcome of most interest, survival when challenged with a dose of anthrax that is 500 times a normally lethal dose to the unvaccinated. The human volunteers were not asked to submit to this exposure.

However, if the macaques are to be challenged but the humans are not, how do we learn about the relative success of different vaccination regimens in humans? The key thought is to use the immunogenicity measurements in humans as 'biomarkers' or 'surrogate outcomes' for survival when challenged, relying on the immunogenicity–survival relationship in macaques to calibrate how much survival to expect with a specific level of observed immunogenicity in humans. In fact, immunogenicity is very complex, as is reactogenicity; in these trials there are many hundred (in fact, apparently nearly 2000) measurements taken across time for each person.

To focus on fundamental issues of direct and indirect causal effects (e.g. is there any direct effect of vaccination after controlling for immunogenicity?), I simplify both trials: first, there are only two levels of vaccination, in the human trial, $W^* = 1$ for low, $W^* = 2$ for high; in the macaque trial, $W = 1$ for low, $W = 2$ for high; and secondly, the planned 'surrogate outcome' or 'biomarker' (= immunogenicity score) is scalar, S^* in the human trial, S in the macaque trial. Survival to challenge is Y (1 = alive/0 = dead), but is only available in the macaque trial. Display 2 summarizes the notation using the potential outcomes framework, where the unobserved survival in humans when challenged is labelled Y^*.

The ultimate objective is to know what levels of vaccine with humans provide protection. That is, in this simplified example, we want to know $Y^*(1)$, survival when exposed to $W^* = 1$, and $Y^*(2)$, survival when exposed to $W^* = 2$. We would also like to know, among those vaccinated with $W^* = 1$, how survival depends on $S^*(1)$ and covariates such as age and sex,

Display 2. *Simplified set-up with two levels of vaccination: low versus high*

Human trial	Macaque trial
W^* = Treatment assignment, 1 or 2	W = Treatment assignment, 1 or 2
$S^*(1)$ = Measured biomarkers when $W^* = 1$	$S(1)$ = Measured biomarkers when $W = 1$
$S^*(2)$ = Measured biomarkers when $W^* = 2$	$S(2)$ = Measured biomarkers when $W = 2$
$Y^*(1)$ = Survival after challenge when $W^* = 1$	$Y(1)$ = Survival after challenge when $W = 1$
$Y^*(2)$ = Survival after challenge when $W^* = 2$	$Y(2)$ = Survival after challenge when $W = 2$

and similarly with the $W^* = 2$ humans, so that we could, for example, revaccinate an individual when protection probability as indicated by S^* is too low.

How do we plan to satisfy these objectives without challenging humans? There are really four steps to this process, where the last is critical and seems to require the use of potential outcomes and principal stratification (Frangakis & Rubin, 2002) to make clear.

Step 1: Predict $S^*(1)$ from humans assigned $W^* = 1$
Predict $S^*(2)$ from humans assigned $W^* = 2$
Step 2: Predict $S(1)$ from macaques assigned $W = 1$
Predict $S(2)$ from macaques assigned $W = 2$

Both of these steps are conceptually straightforward where the predictive model can also include covariates; these steps are also practically straightforward in the absence of unintended missing data, which are expected to be a major nuisance in the human trial but are not discussed here.

Step 3: Predict $Y(1)$ from $S(1)$ from macaques assigned $W = 1$
Predict $Y(2)$ from $S(2)$ from macaques assigned $W = 2$

Again, step 3 is conceptually straightforward, and may include covariates, such as sex and baseline measurements.

Step 4: Use the results of step 3 as if they applied to humans, thereby implicitly assuming that $W^* = W$, $S^* = S$ and $Y^* = Y$, and use the prediction model found in step 3 to:
Predict unobserved $Y^*(1)$ from $S^*(1)$ from humans assigned $W^* = 1$
Predict unobserved $Y^*(2)$ from $S^*(2)$ from humans assigned $W^* = 2$

Step 4 as stated is deceivingly simple: Y^* and Y have the same unambiguous meaning in humans and macaques (alive versus dead), and S^* and S have very similar meanings (e.g. antibody density), but how is a dose of vaccine, W^*, in a 100 kg human to be equated to a dose, W, in a 6 kg macaque?

4. Using immunogenicity as a surrogate/biomarker

There is no clear reason for, *a priori*, equating a particular dose of vaccine, W^*, in a human with a particular dose W in a macaque. One might think that the 'equivalent' dose should be per kg of body weight, but the medical researchers who were most familiar with the past studies felt that doses much closer to the same absolute dose were appropriate (e.g. low for humans = 1 cm^3, low for macaques = 1 cm^3) because of threshold effects. How should we argue in step 4?

The hope is, vaguely, that S^* and S 'have the same meaning' for survival in the sense that S is a 'surrogate/biomarker' for survival under challenge in macaques, and S^* is similarly a 'surrogate/biomarker' in humans, so that there is no 'direct' effect of W on Y given S, nor of W^* on Y^* given S^*. The only place to look for support for this claim is in the macaques where W, $S(W)$ and $Y(W)$ are all measured.

So in some sense, we are now forced to look in the macaques for no 'direct effects' of W on Y; all effects of W on Y should be through S. Or, the 'causal pathway' of W to Y should be through S. This sounds good, but how do we translate this into data analysis? The seductive picture

$$W \rightarrow S \rightarrow Y,$$

with no arrow from W to Y, seems to suggest looking in the macaques for evidence in the observed data (W_{obs}, S_{obs}, Y_{obs}) that, given S_{obs}, there is no relationship between Y_{obs} and W_{obs}. But reality may not be that simple.

Display 3 depicts the issue via potential outcomes and principal stratification. The principal strata in this case are defined by the pair of values of the proposed surrogate $(S(1),S(2))$, where for simplicity we take S to be binary, either L = low immunogenicity or H = high immunogenicity. Furthermore, we assume three principal strata. Stratum 1 includes those with $S(1) = S(2) = L$, representing the collection of macaques whose vaccination, whether at a high or low level, would result in low immunogenicity – the LL group. Stratum 2 includes those with $S(1) = L$ and $S(2) = H$, representing the collection of macaques who, if exposed to the low level vaccination would have low immunogenicity, but if exposed to the high level of vaccination would have high immunogenicity – the LH group. Stratum 3 includes those with $S(1) = S(2) = H$, representing the collection of macaques whose immunogenicity would be high whether the vaccination was at a high level or a low level – the HH group. The fourth theoretically possible principal stratum, the collection of macaques whose immunogenicity would be low with high-level vaccination and high with low-level vaccination, makes no scientific sense, and therefore we assume the HL group to be void.

Consider the hypothetical potential outcome data displayed in the left half of Display 3. The top and bottom displays depict two situations. Both situations reflect beneficial effects of high level vaccination in the principal stratum where vaccination affects immunogenicity, that is, in the LH group, where vaccination increases the survival rate from 40 to 60 per cent in both situations. The two situations differ in that the top one reflects a benefit of high- versus low-level vaccination within each principal stratum, whereas the bottom display reflects no effect of high- versus low-level vaccination for those for whom vaccination does not affect immunogenicity.

In the top situation, there appears to be a 'direct' causal effect of vaccine on survival revealed by the effect of W on $Y(W)$ in the two principal strata with $S(1) = S(2)$. But what language is appropriate to describe this situation? Is the survival effect 'not mediated' by immunogenicity? Is immunogenicity 'not on the causal pathway' from vaccination to survival?

In the bottom situation, there appears to be no 'direct' effect of vaccination, all effect being 'indirect', 'mediated' by immunogenicity, because if vaccination cannot change immunogenicity, it cannot affect survival probability. Is it appropriate to say in this case that immunogenicity is 'on the causal pathway' from vaccine to survival? Is immunogenicity a biomarker in this situation but not in the first?

Instead of answering these semantic questions, we consider the observed data for these two situations in randomized experiments, and the resulting inference for 'direct' and 'indirect' effects.

Display 3. *Two examples of the presence and absence of direct effects*

Principal stratum (equal sized)	Potential outcomes: Surrogates $S(1)$	$S(2)$	Survival % $Y(1)$	$Y(2)$	Observed data $(S_{obs}, \bar{Y}_{obs})$ given treatment assignment: $W_{obs} = 1$	$W_{obs} = 2$
a. Case where there is a direct causal effect of W on Y given $S(1)$, $S(2)$, but W_{obs} and Y_{obs} are conditionally independent given S_{obs}						
1.	L	L	0	20	L, 20	L, 20
2.	L	H	40	60		H, 80
3.	H	H	80	100	H, 80	
b. Case where there is no direct causal effect of W on Y given $S(1)$, $S(2)$, but W_{obs} and Y_{obs} are conditionally dependent given S_{obs}						
1.	L	L	0	0	L, 20	L, 0
2.	L	H	40	60		H, 70
3.	H	H	80	80	H, 80	

5. The observed data in the hypothetical example

For simplicity, assume an equal number of macaques in each principal stratum in both situations depicted in Display 3. This assumption is entirely innocuous in this case with three principal strata, because the proportions in each situation can always be estimated from observed data, W_{obs}, S_{obs}, as we show in section 6. In randomized experiments with half given $W = 1$ and half given $W = 2$, we would expect to see the observed data shown in the right half of Display 3, where the boxes indicate groups with the same value of W_{obs} and S_{obs}, and the corresponding proportion surviving ($\bar{Y}_{obs}$) is indicated.

In the top situation in Display 3 we *appear* to have strong evidence that all the effect of the vaccine is via immunogenicity because conditionally given S_{obs}, Y_{obs} and W_{obs} are independent. That is, for a given level of observed immunogenicity (S_{obs}), the protection (Y_{obs}) is the same whether high- or low-level vaccination (W_{obs}) took place. To be explicit, when $S_{obs} = L$, $Y_{obs} = 20$ per cent whether $W_{obs} = 1$ or 2; and when $S_{obs} = H$, $Y_{obs} = 80$ per cent whether $W_{obs} = 1$ or 2. But we know, from an examination of the left half of Display 3 using the unobserved potential outcomes, that these words do not accurately describe reality, where, as far as we can tell, all macaques benefit equally from a high- versus low-level vaccination, no matter what their attained immunogenicity.

Now examine the bottom situation. Here, vaccination *appears* from the observed data to have a 'direct' effect on survival that is 'mediated' by the immunogenicity level: when observed immunogenicity is low, high-level versus low-level vaccination appears to reduce survival from 20 to 0 per cent, whereas when observed immunogenicity is high, high-level vaccination appears to reduce survival from 80 to 70 per cent; so does high-level vaccination reduce survival for all? But again, this is an inaccurate conclusion, as is clear from the left half of the display involving the potential outcomes: unless immunogenicity is altered, high versus low vaccination has no effect on survival, and it is beneficial for one-third of the group and has no effect on the other two-thirds. Of course, the analysis of Y_{obs} ignoring S_{obs} leads to a valid conclusion about the causal effect of W on Y because of the randomization of W.

I will leave it to others to produce the correct graphical displays for these two situations, but clearly, the naive one displaying the conditional independence between W_{obs} and Y_{obs} given S_{obs} is seductive but leads to the incorrect conclusion. How then should we deal with the observed data in these situations?

My conceptual approach is to think about how to multiply impute the missing potential outcomes, both immunogenicity and survival, and thereby multiply impute the principal strata. In order to define a predictive distribution from which to draw the imputations for the missing observations, assumptions must be made, and they must be made explicitly. I regard this as an advantage, not a disadvantage, of this approach relative to the seductive graphical approach.

6. Multiply imputing missing potential outcomes

For over three decades, I have believed that all problems of causal inference should be viewed as problems of missing data: the potential outcomes under the not-received treatment are the missing data. A straightforward and valid way to think about missing data is to think about how to multiply impute them.

Multiple imputation is distributional prediction. For each imputed data set, a correct answer is obtained by simple calculations. Repeating the imputations over and over again not only reveals the typical (e.g. average) correct answer, but also the uncertainty about this answer. As we humans seem generally inept at dealing with uncertainty directly, this technique can be extremely helpful because the uncertainty is handled automatically. Assumptions made

when creating the imputations are crucial, but the use of multiple imputation does allow easy sensitivity assessments of the results to various assumptions.

Only one set of potential outcomes, in our case either $S(1), Y(1)$ or $S(2), Y(2)$, is ever observed on the same unit (the 'fundamental problem of causal inference', Holland, 1986). When there is no desire to draw causal inferences conditional on potential outcomes, this lack of data on the joint distribution of potential outcomes under different treatment assignments creates only trivial practical problems, although it creates interesting theoretical issues, as discussed in Neyman (1923) from the frequentist perspective, and Rubin (1978, 1990) from the Bayesian perspective. In our case, however, there is a desire to condition on the joint values of the immunogenicity potential outcomes, $S(1)$, $S(2)$, and so the specification of its unobservable joint distribution becomes critical.

If we had pretreatment covariates X measured for everyone, these would imply that it is the joint *conditional* distribution of $S(1)$, $S(2)$ given X that is unobservable, and this can substantially limit the extent of the relationship between $S(1)$ and $S(2)$ that is unobserved, depending on the strength of the covariate X in predicting the margins of $S(1)$ and $S(2)$. This point has been made often in the missing data literature (e.g. Rubin & Thayer, 1978), and in the causal inference literature generally (e.g. Rubin, 1978), as well as in the context of principal stratification (Zhang, 2002; Zhang & Rubin, 2004).

Within levels of X, we can place restrictions on the possible values of the potential outcomes, either using Bayesian prior distributions, as in Imbens & Rubin (1997) and Hirano *et al.* (2000), or absolute restrictions, as in traditional econometric instrumental variables models, as discussed in Angrist *et al.* (1996). The latter are often called 'exclusion restrictions' in economics because they exclude certain values of potential outcomes as impossible. We have already done this here, when we excluded the fourth type of principal stratum, by declaring the HL stratum void. This assumption allows us to estimate the population proportions in each of the three remaining principal strata: the individuals with $W_{\text{obs}} = 1$, $S_{\text{obs}} = H$ are observed to be a third of those randomized to $W = 1$, and can only belong to the HH stratum; analogously the individuals with $W_{\text{obs}} = 2$, $S_{\text{obs}} = L$ are observed to be a third of those randomized to $W = 2$, and can only belong to the LL stratum; hence in large samples, we know, in both cases of Display 3, that the principal stratum proportions are 1/3, 1/3, 1/3. Which additional restrictions are plausible depends critically both on the science of the situation and the chosen design (the assignment mechanism). Such restrictions can be quite beneficial by sharpening the implied distribution of the missing potential outcomes that are to be imputed.

7. Some possible assumptions on potential outcomes

In some cases, additional but plausible, exclusion restrictions allow full identification in the sense of implying unique large sample point estimates of causal effects within each principal stratum. For example, suppose, in our problem, that in addition to assuming no HL group, we assume that if high versus low vaccination has no effect on immunogenicity, then it has no effect on survival, as in the bottom half of Display 3: that is, if $S(1) = S(2)$, then $Y(1) = Y(2)$. Under this assumption, the only non-null causal effect of W on Y occurs in the LH principal stratum, which we know (in large samples) comprises one-third of the population. A simple algebraic argument (Angrist *et al.*, 1996) shows that the causal effect of W on Y in the LH group is then given by the overall causal effect of W on Y (found by ignoring S altogether) divided by the proportion in the LH group. Such an assumption will greatly sharpen the multiple imputations that can be created for the missing potential outcomes (as illustrated in Imbens & Rubin, 1997; Hirano *et al.*, 2000).

A weaker assumption, but highly plausible in our case, is to allow causal effects of vaccination in all principal strata, but require them to be non-negative. That is, even when $S(1) = S(2)$, we require $Y(1) \geq Y(2)$, so that high- versus low-level vaccination cannot reduce survival. The top part of Display 3 satisfies this assumption, although not the previous one that disallows a causal effect of W on Y if there is no causal effect of W on S. Under the weaker assumption, generally we cannot get full identification in the sense of the previous paragraph, but we can obtain large-sample bounds on all causal effects, and can still limit the range of multiple imputations. Without any assumption other than no HL group, bounds can be obtained but these are typically quite wide. To develop this topic carefully requires extra notation, and so is beyond the scope of this article; see Zhang & Rubin (2004) for analogous calculations in the context of 'censoring due to death'.

Alternatively, suppose we assumed that within each level of vaccination (still randomized with probability 1/2), immunogenicity was 'essentially randomized' with probability $W/3$ to be high and probability $1 - W/3$ to be low. Then we would view the data as arising from four randomized treatment groups (W, S) with relative sizes $(1, L) = 1/3$; $(1, H) = 1/6$; $(2, L) = 1/6$; $(2, H) = 1/3$. Then Display 3 using principal strata defined by S would not be generally appropriate because S is no longer an outcome of W but is randomized within levels of W. That is, the assumption is that for the macaques randomized to $W = 1$, those with $S_{\text{obs}} = 1$ are only randomly different from those with $S_{\text{obs}} = 2$, and similarly for the macaques randomized to $W = 2$. Covariates can help to make this assumption more believable: if within levels of W, the covariates predict S_{obs} very well, we may well believe any residual variability in S has been essentially randomized. That is, for each macaque with $S_{\text{obs}} = L$, S_{obs} could have been H if the biased coin assigning S had flipped the other way, and analogously for the macaques with $S_{\text{obs}} = H$. With covariates, the bias of this hypothetical coin can depend not only on W but also on the covariates, thereby implying an ignorable treatment assignment mechanism with $Y(W, S)$ being the only outcome variable, $W = 1, 2$ and $S = L, H$. In this case, the multiple imputation of missing potential outcomes is standard because we simply have a four-treatment randomized experiment with covariates.

Thus, we see the critical role of covariates and extra assumptions in reducing variability in the imputed values of the missing potential outcomes and returning us to the inferential situation where we are comfortable. I find the potential outcomes framework implemented via multiple imputation to be the most revealing way to address these issues, and I hope that this presentation has provoked stimulating and clarifying discussion.

References

Adams, P., Hurd, M. D., McFadden, D., Merrill, A. & Ribeiro, T. (2003). Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status. *J. Econometrics* **112**, 3–56.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion and rejoinder). *J. Amer. Statist. Assoc.* **91**, 444–472

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd, London.

Frangakis, C. & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86**, 366–379.

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

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica* **12** (Suppl.), 1–115.

Hirano, K., Imbens, G., Rubin, D. B. & Zhao, X. H. (2000). Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1**, 69–88.

Holland, P. (1986). Statistics and causal inference (with discussion and rejoinder). *J. Amer. Statist. Assoc.* **81**, 945–970.

Imbens, G. W. & Rubin, D. B. (1995). Discussion of 'Causal diagrams for empirical research' by J. Pearl. *Biometrika* **82**, 694–695.

Imbens, G. W. & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25**, 305–327.
Mealli, F. & Rubin, D. B. (2003). Assumptions allowing the estimation of direct causal effects: commentary on 'Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status' by Adams *et al*. *J. Econometrics* **112**, 79–87.
Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles (with discussion). Section 9 (translated). *Statist. Sci.* **5**, 465–480.
Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82**, 669–688.
Pearl, J. (2001). Direct and indirect effects. In *Proceedings of 7th Conference on Uncertainty in artificial intelligence*, (eds J. S. Breese & D. Koller), 411–420. Morgan Kaufmann, San Francisco, CA.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.
Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *J. Educ. Statist.* **2**, 1–26 (Printer's correction note 3, p. 384).
Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **7**, 34–58.
Rubin, D. B. (1980). Discussion of 'Randomization analysis of experimental data in the Fisher randomization test' by Basu. *J. Amer. Statist. Assoc.* **75**, 591–593.
Rubin, D. B. (1990). Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.* **5**, 472–480.
Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with noncompliance (with discussion 387–389). *Statist. Med.* **17**, 371–385
Rubin, D. B. (2000). The utility of counterfactuals for causal inference. Comment on A.P. Dawid, 'Causal inference without counterfactuals'. *J. Amer. Statist. Assoc.* **95**, 435–438.
Rubin, D. B. & Thayer, D. T. (1978). Relating tests given to different samples. *Psychometrika* **43**, 3–10.
Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13**, 238–241.
Tinbergen, J. (1930). Determination and interpretation of supply curves: an example, *Zeitschrift für Nationalökonomie*. Reprinted in: *The foundations of econometrics* (eds D. Hendry & M. Morgan), 233–245. Cambridge University Press, Cambridge, UK.
Zhang, J. (2002) *Causal inference with principal stratification: some theory and application*. PhD Thesis, Department of Statistics, Harvard University.
Zhang, J. & Rubin, D. B. (2004). Censoring due to death via principal stratification. To appear in the *J. Educ. Behav. Statist.*

Received November 2002, in final form November 2003

Donald B. Rubin, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA.
E-mail: rubin@stat.harvard.edu