Causal Interpretations Conditional on Surviving

G. Rohwer (July 2014)

Abstract. The article discusses how to think of the causal effect of a treatment X, realized at age t, on an outcome variable Y, whose values only exist if an individual survives at least until $t+\delta$. As a framework, the article uses stochastic potential outcomes defined for generic individuals which are defined by values of variables, without using an identifying name. The article conceives of 'the cause' of Y as a generating process that begins with the treatment but temporally extends until, possibly, the variable Y takes a particular value. In contrast to Rubin's proposal to condition on a 'principal stratum' of 'potential survivors', the article argues that effects defined conditional on actual surviving can well be given a causal interpretation.

Keywords: Truncation by death, principal stratification, stochastic potential outcomes, causal interpretations.

1 Introduction

I refer to a population of individuals, $\mathcal{U}$. The variable X denotes a treatment, realized at age t, which can take two possible values (0 or 1). Z is a vector of pre-treatment variables characterizing the individuals in $\mathcal{U}$. The interest concerns effects of X on a variable Y, taking values at $t+\delta$ in the domain $\mathcal{Y}$. However, a value of Y only comes into being if an individual survives at least until $t+\delta$. This is recorded by a binary variable D, with $D=1$ if the individual died between t and $t+\delta$, and otherwise $D=0$. I consider the question of how to conceptualize a causal effect of X on Y that takes into account that values of Y only exist conditional on $D=0$.

Since $D=0$ is a necessary condition for values of Y to exist, it is not a mediator variable in the usual sense, which presupposes that Y is defined for all possible values of the mediator variable. So it would be misleading

to refer to a diagram

$$\begin{array}{ccc} (Z,X) & \longrightarrow & Y \\ & \searrow \quad \nearrow & \\ & D & \end{array} \tag{1}$$

because, if $D = 0$, there cannot be an arrow leading from (Z, X) to Y. The distinction between a direct and an indirect effect of X is therefore of no use in the present application.

In order to cope with this difficulty I propose to think of outcomes as processes which consist in generating a value of D and, if $D = 0$, also a value of Y. This will be further discussed in Section 2. In Section 3 I introduce a notion of 'potential outcomes' which considers such outcomes as values of random variables. In Section 4 I consider the question of how to understand effects of X on Y conditional on $D = 0$. I criticize Rubin's proposal that one should condition on variables intended to represent 'potential surviving', instead of D. In order to suggest a causal interpretation, I consider (Z, X, D) as a process which, possibly, generates values of Y. I then discuss how to understand a corresponding causal claim. The paper ends with a brief conclusion.

2 Outcomes as processes

I propose to think of effects of X by referring to processes which are initiated by this variable's taking a particular value. These processes take place in a context given by a value of Z and consist in two steps: the variable D takes a particular value, and then, if this is zero, also the variable Y takes a particular value.

In order to represent these processes, I use an extended version of Y, denoted by Y^*, which can take values in $\mathcal{Y}^* := \mathcal{Y} \cup \{*\}$. Restricted to $\mathcal{Y}$, $Y^* = Y$; in addition, $Y^* = *$ means that a value of Y does not exist. For each individual $i \in \mathcal{U}$, the observed outcome can be described as a process (d_i, y_i^*), where d_i and y_i^* are i's values of D and Y^*, respectively.

One can now start from a joint distribution of D and Y^*, conditional on values of X and Z. Respecting the temporal ordering, this joint distri-

bution can be expressed as

$$\begin{aligned}\Pr(D=d, Y^*=y^* \mid X=x, Z=z) = \\ \Pr(D=d \mid X=x, Z=z)\Pr(Y^*=y^* \mid D=d, X=x, Z=z)\end{aligned} \tag{2}$$

Since $\Pr(Y^*=* \mid D=d, X=x, Z=z) = d$ is known in advance, the joint distribution can be derived from knowing

$$\Pr(D=0 \mid X=x, Z=z) \tag{3}$$

and

$$\Pr(Y=y \mid D=0, X=x, Z=z) := \Pr(Y^*=y \mid D=0, X=x, Z=z) \tag{4}$$

Note that this is a definition of a conditional distribution for Y.

3 Potential outcomes

In order to discuss causal interpretations it has been proposed to refer, for each individual $i \in \mathcal{U}$, to 'potential outcomes' resulting from $X=0$ or $X=1$, respectively (e.g. Rubin 2000; 2006). Here I follow this proposal but conceive of potential outcomes as random variables.

The most often used notion of potential outcomes is deterministic and, for the present application, can be described by the diagram

$$\begin{array}{ccc} I & \longrightarrow & (D, Y^*) \\ & \nearrow & \\ X & & \end{array} \tag{5}$$

where I is a variable whose values are identifiers of the individuals in $\mathcal{U}$. It is assumed that the relationship is deterministic entailing the existence of a function

$$(d_i, y_i^*) = h(i, x_i) \tag{6}$$

Note that an explicit reference to Z is not required because $I = i$ deterministically entails $Z = z_i$.

In this paper I use a stochastic notion of potential outcomes. Instead of a reference to identifiable individuals, the basic reference is to generic

individuals which are defined by values of variables, without using an identifying name. (5) is replaced by

$$\begin{array}{ccc} Z & \longrightarrow & (D, Y^*) \\ & \nearrow & \\ X & & \end{array} \tag{7}$$

and the relationship is described by a conditional probability distribution as already given in (2). Following this approach, potential outcomes can be defined as random variables. For example, one can define a random variable for potential surviving, $D_z(x)$, whose distribution is defined by

$$\Pr(D_z(x) = d) := \Pr(D = d \,|\, X = x, Z = z) \tag{8}$$

Note the distinction between *potential* outcomes, which are random variables, and *possible* outcomes, which are values of such random variables. Each potential outcome can be described by a probability distribution for a set of possible outcomes. However, since potential outcomes are distinguished by values of variables, one cannot define joint distributions of potential outcomes. For example, since X cannot simultaneously take the values 0 and 1, there is no joint distribution of $D_z(0)$ and $D_z(1)$.

I now follow the idea to think of causal effects as quantities informing about aspects of a comparison between two potential outcomes for 'the same individual' (e.g. Rubin, 2005). In a deterministic approach, one would have to compare potential outcomes for an identifiable individual, which is impossible. Following the stochastic approach, one has to compare potential outcomes for generic individuals defined by values of Z, and this is possible. In the present application, the comparison concerns the potential outcomes

$$(D, Y^*)_z(0) \;\text{ and }\; (D, Y^*)_z(1)$$

which result from the two possible treatments. A complete characterization would require to describe the corresponding probability distributions separately. For simplification, one can separately refer to the two terms on the right-hand side of (2). In order to characterize a difference between the two versions of the first term one could use the quantity

$$\Delta^D(z) := \Pr(D = 0 \,|\, X = 1, Z = z) - \Pr(D = 0 \,|\, X = 0, Z = z) \tag{9}$$

describing the effect of X on the probability of surviving. With respect to the second term on the right-hand side of (2) one can use (4), and a further simplification is then possible by only considering expectations of Y. Some information is then given by

$$\Delta^Y(z) := \qquad (10)$$
$$\mathrm{E}(Y \mid D=0, X=1, Z=z) - \mathrm{E}(Y \mid D=0, X=0, Z=z)$$

Of course, even both quantities together only provide partial information about the whole effect that shows up in the dependence of the joint distribution of D and Y^* on X, given the context $Z = z$.

4 Interpreting effects on Y

While the meaning of $\Delta^D(z)$ is easily understood, it is not obvious how to understand $\Delta^Y(z)$. In this section I first consider an argument, made in particular by Rubin (2000; 2006), that $\Delta^Y(z)$ should not be understood as a causal effect. I then propose a causal interpretation.

Principal stratification

Rubin's main critique is that, if one conditions on $D = 0$, one does not compare potential outcomes for a common set of individuals (Rubin, 2006). The argument is based on assuming the existence of variables $\tilde{D}^x$, for $x=0,1$, having values

$$\tilde{d}_i^x := \begin{cases} 0 & \text{if individual } i \text{ would survive at least until } t+\delta \text{ if } x_i = x \\ 1 & \text{otherwise} \end{cases}$$

Rubin claims that to think of a causal effect of X on Y is only sensible for individuals belonging to a 'principal stratum' defined by $\mathcal{U}_c := \{i \in \mathcal{U} \mid \tilde{d}_i^0 = \tilde{d}_i^1 = 0\}$, that is, individuals who would survive, and therefore have values of Y, regardless of their treatment. He then suggests that a causal effect should be defined by

$$\Delta_c^Y(z) := \mathrm{E}(Y \mid \tilde{D}^0=\tilde{D}^1=0, X=1, Z=z) - \qquad (11)$$
$$\mathrm{E}(Y \mid \tilde{D}^0=\tilde{D}^1=0, X=0, Z=z)$$

that is, restricted to the 'principal stratum' $\mathcal{U}_c$.

There are two difficulties. First, values of $\tilde{D}^x$ cannot be observed, and one would need questionable assumptions for the identification of membership in $\mathcal{U}_c$ (e.g., Zhang and Rubin, 2003; Egleston et al., 2007; Lee et al., 2010; Ding et al., 2011). Therefore, quantities defined conditional on these variables do not have clear empirical applications. Second, these variables presuppose that an individual's survival status at $t + \delta$ is determined already at time t when the treatment is generated. In fact, since Rubin requires that a random assignment of treatments makes X and $\tilde{D}^x$ independent (conditional on values of Z), one has to assume that an individual's survival status at $t + \delta$ is determined already before the time when the treatment, on which survival depends, is generated. Assuming the existence of these variables is therefore not compatible with the view that values of D result from a contingent process following the treatment. (For additional critique, and further references, see Dawid and Didelez, 2012.)

Following the stochastic conception of potential outcomes introduced in Section 3, variables like $\tilde{D}^x$ cannot be defined. One would have to use stochastic potential outcomes, $D_z(x)$, whose distributions are defined by (8). The expression $D_z(x) = d$ is to be used for referring to a possible state of affairs whose realization can be assigned a probability. On the other hand, values of variables to be used as conditions in a conditional probability statement must be understood as (hypothetically) realized facts. Using $D_z(x) = 0$ as a condition has therefore no clear meaning. It might be interpreted as meaning that there is a positive probability for surviving; but then the condition will be true for all members of $\mathcal{U}$ (presupposing treatments that will not deterministically entail death).

Causes as generating processes

I now return to the question of how to understand $\Delta^Y(z)$. Interpretations require a reference to the process generating values of Y. This process begins with the generation of a value of X, in a context given by a value of Z, and includes all further events which possibly happen afterwards until $t + \delta$ and are causally relevant for the coming into existence of values of Y. The simple set-up considered here takes into account just one of these

events, namely surviving. So one can use the diagram

$$(Z, X, D) \longrightarrow Y^* \tag{12}$$

The cause of Y^* is conceptualized as a generating process. Potential outcomes will be denoted by $Y_z^*(x, d)$, having distributions

$$\Pr(Y_z^*(x, d) = y) := \Pr(Y^* = y \mid D = d, X = x, Z = z) \tag{13}$$

There are now four potential outcomes for each generic individual, depending on x and d.

$\Delta^Y(z)$ can be understood as characterizing a comparison of two of these potential outcomes, namely $Y_z^*(0, 0)$ and $Y_z^*(1, 0)$. So it agrees to the demand that a causal effect should be defined by a reference to two potential outcomes for the same generic individual.

Further demands are less clear. Adherents to a potential outcomes approach often require that variables representing presumed causes should be independent of the potential outcomes (e.g. Rubin, 2008). Our presumed cause is the generating process (Z, X, D), and since

$$\Pr(Y_z^*(x, d) = y, X = x', D = d', Z = z)$$

equals $\Pr(Y_z^*(x, d) = y)$ if $x = x'$ and $d = d'$, and is zero (or undefined) otherwise, this requirement of independence is trivially fulfilled (or senseless) and therefore of no value when potential outcomes are conceptualized as random variables. (Even in the case of deterministically conceptualized potential outcomes, the meaning of the independence requirement is obscure. Greenland, Robins and Pearl (1999, p. 42) have suggested the following interpretation (where $y_{i0}, \ldots, y_{iK}$ is their notation for potential outcomes): "[T]he analyst must be prepared to treat $y_{i0}, \ldots, y_{iK}$ as parameters unaffected by treatment assignment. Treatment assignment only determines which of these $K + 1$ parameters we observe [that is, the realization of Y_i (Rubin, 1974, 1978, 1991)]; the other K parameters remain latent traits of individual i." It seems not possible to reconcile this interpretation with the ordinary understanding that treatments are events contributing to the generation of new facts.)

Another idea is that the variable representing the treatment should be

independent of all other variables which are causally relevant for the outcome. This requirement is, however, ambiguous when the treatment is considered as an integral part of a generating process. One could require that the variable representing the treatment is independent of all pre-treatment variables, but one cannot demand this for variables which temporally follow the treatment in the generating process. In the present application, since X is assumed to be causally relevant for D, these variables cannot be independent (regardless of whether values of X are randomly assigned). It seems obvious, however, that this is not required for justifying the claim that the generating process, (Z, X, D), is causally relevant for Y^*.

Almost always there are at least some components of Z which are causally relevant also for D. It follows that the generating subprocess, (X, D), cannot be independent of Z, even if values of X are randomly assigned. However, since Z is explicitly considered as an essential part of the complete generating process, (Z, X, D), this dependence is no hindrance to a causal interpretation of the complete process, or of the subprocess conditional on values of Z.

Unobserved covariates

There is, however, a further difficulty. The cited requirement refers to 'all other variables' which are causally relevant for the outcome, and one most often has to assume that some of these variables are not observed and therefore not part of Z. So let V denote a vector of further unobserved pre-treatment variables assumed to be causally relevant for D and Y^*. As in (10), one can condition on values of Z and V resulting in conditional effects $\Delta^Y(z, v)$. These effects cannot be estimated, but one can think of a mean effect

$$\bar{\Delta}^Y(z) := \sum\nolimits_v \Delta^Y(z, v) \Pr(V = v) \tag{14}$$

where $\Pr(V = v)$ denotes an unknown distribution, realized before t. Note that this mean effect depends on the presupposed distribution of V if $\Delta^Y(z, v)$ depends on v.

In general, $\bar{\Delta}^Y(z) \neq \Delta^Y(z)$. This is an immediate consequence of the fact that V is assumed to be causally relevant for D and therefore cannot be independent of (Z, X, D). Note that this cannot be avoided by a random

assignment of treatments since even then V and D will still be dependent. So the question arises how to interpret the quantity $\Delta^Y(z)$ which can be estimated.

One possibility is to consider $\Delta^Y(z)$ as a biased estimate of $\bar{\Delta}^Y(z)$. At first sight, this seems to follow from the requirement that 'average causal effects' should be conceptualized as averages of 'individual causal effects' (e.g. Rubin, 2005; Little and Rubin, 2000). This requirement is ambiguous, however, because it depends on a reference to 'individuals'.

In our framework, 'individuals' must be considered as generic individuals. Having introduced the variables Z and V, generic individuals can be distinguished according to values of these variables. An average of 'individual effects' is then given, for example, by $\bar{\Delta}^Y(z)$. However, if V is not observed, it is not possible to use values of V to distinguish between individuals. Instead, generic individuals can only be defined by a reference to Z, and all individuals with $Z = z$ are considered as exchangeably representing the generic individual z. Consequently, one can interpret $\Delta^Y(z)$ either as a generic effect or, equivalently, as an average of individual effects which are identical for all individuals with $Z = z$.

The argument is not meant to say that one can simply ignore unobserved causally relevant covariates; it is meant to suggest a somewhat different view of the problem. Note again that randomization of X does not make $\Delta^Y(z)$ an unbiased estimate of $\bar{\Delta}^Y(z)$. In fact, if V is assumed to represent *all* variables which, in addition to Z, could be causally relevant for D and Y, there is no way to estimate $\bar{\Delta}^Y(z)$, and this quantity is to be considered as a pure theoretical fiction. Characterizing $\Delta^Y(z)$ as a biased estimate of $\bar{\Delta}^Y(z)$ is therefore not a useful critique. Instead, one should look for identifiable particular covariates which could be made observable and then included in order to establish more detailed causal insights.

5 Conclusion

I have considered how to think of the causal effect of a treatment X, realized at age t, on an outcome variable Y, whose values only exist if an individual survives at least until $t + \delta$. As a framework, I have used stochastic potential outcomes defined for generic individuals which are

defined by values of variables, without using an identifying name. I have discussed two complementary approaches. The first approach is based on conceiving of a treatment, $X = x$, as the starting point of a process which first generates surviving, or death, and, if surviving, also a value of Y. This allows one to define measures of effect which characterize a comparison of such processes.

As a complementary approach, I have conceived of 'the cause' of Y as a generating process that begins with the treatment but temporally extends until, possibly, the variable Y takes a particular value. In contrast to Rubin's proposal to condition on a 'principal stratum' of 'potential survivors', I have argued that effects defined conditional on actual surviving can well be given a causal interpretation.

As part of the argument, I have discussed how to think of unobserved covariates which are, presumably, causally relevant both for surviving and Y. I have argued that, although marginal effects can then be distinguished from averages of conditional effects, also marginal effects can be considered as causal effects for generic individuals defined on the basis of observable variables.

Note that the reference to generic, instead of identifiable, individuals is a consequence of a probabilistic approach to causal effects. Since a treatment can be applied at most once to each particular individual, it would not make sense to use probability distributions of outcomes conditional on $I = i$ (instead of values of proper variables). This is no hindrance to probabilistically predict causal effects for a particular individual, say i^*. However, such predictions must be based on knowing i^*'s value of the variables on which the causal effect depends. In the present application, knowing the individual's value of Z, say z_{i^*}, one could use $\Delta^D(z_{i^*})$ and $\Delta^Y(z_{i^*})$. In order to criticize such predictions one would not only need to know more nuanced causal effects, as e.g. $\Delta^D(z, v)$ and $\Delta^Y(z, v)$; but also i^*'s values of Z and V.

References

Dawid, P. and Didelez, V. 2012. "'Imagine a Can Opener'—The Magic of Principal Stratum Analysis." *The International Journal of Biostatistics* 8(1),

Article 19.

Ding, P., Geng, Z., Yan, W. and Zhou, X-H. 2011. "Identifiability and Estimation of Causal Effects by Principal Stratification with Outcomes Truncated by Death." *Journal of the American Statistical Association* 106, 1578–91.

Egleston, B. L., Scharfstein, D. O., Freeman, E. E. and West, S. K. 2007. "Causal Inference for Non-mortality Outcomes in the Presence of Death." *Biostatistics* 8, 526–545.

Greenland, S., Robins, J. M. and Pearl, J. 1999. "Confounding and Collapsibility in Causal Inference." *Statistical Science* 14, 29–46.

Lee, K., Daniels, M. J. and Sargent, D. J. 2010. "Causal Effects of Treatments for Informative Missing Data due to Progression/Death." *Journal of the American Statistical Association* 105, 912–929.

Little, R. J. and Rubin, D. B. 2000. "Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches." *Annual Review of Public Health* 21, 121–145.

Rubin, D. B. 2000. "Comment on 'Causal inference without counterfactuals'." *Journal of the American Statistical Association* 95, 435–438.

Rubin, D. B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100, 322–331.

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.

Rubin, D. B. 2008. "Statistical Inference for Causal Effects, With Emphasis on Applications in Epidemiology and Medical Statistics." Pp. 28–63 in *Handbook of Statistics*, Vol. 27, edited by C. R. Rao, J. P. Miller and D. C. Rao. Amsterdam: Elsevier.

Zhang, J. L. and Rubin, D. B. 2003. "Estimation of Causal Effects via Principal Stratification When Some Qutcomes are Truncated by 'Death'." *Journal of Educational and Behavioral Statistics* 28, 353–368.