

Defining and Identifying the Effect of Treatment on the Treated

S. Geneletti A.P.Dawid
Imperial College Cambrigde University

November 6, 2007

1 Introduction

One rôle of labour economics is to evaluate the impact of government initiatives, such as employment and education schemes, on economic indicators such as income, the purpose being to inform the introduction of future schemes and policy changes: Should more adult training programmes be funded? Should education be made compulsory until the age of 18? Evaluating the effect of such policies is far from straightforward, most especially on account of the fundamental problem of *self-selection* (Heckman, 1979). Because of self-selection it is typically unclear whether, and to what extent, changes in economic indicators can be attributed to government programmes where participation is voluntary, since individuals that take part in such programmes (*i.e.*, the self-selected) tend to be more motivated and receive higher incomes, irrespective of participation.

A similar problem emerges in epidemiologic contexts. A recent ruling in the US (Okie, 2006) gives terminal cancer patients the right to be treated with experimental (Phase I) drugs. This means that patients can self-select themselves into the treatment group, without randomisation. Data on response from this group of patients will not yield reliable estimates of the *average causal effect* (ACE) as estimates will be confounded by the patients' attitude and health.

We will use the following two examples, one from labour economics and another from epidemiology, where effect evaluation is hampered by unknown selection criteria, to illustrate aspects of the methodology we develop in this paper.

Example 1 Training programme. As a local government initiative, a mathematics refresher course aimed at adults with no higher education is introduced into a community. After some time, the local government wants to know whether and to what extent the course has had an impact on the income of the participants, as it plans to introduce more refresher courses and make participation in such courses a requirement for job-seekers enrolled in employment schemes. The problem with evaluating the impact of the initiative is that it will be confounded by partially unobserved individual characteristics. Thus estimates of the effect are typically obtained by relying on additional—and generally untestable—assumptions (Heckman, 1979). □

Example 2 Invalid randomisation. Consider an epidemiologic trial where drug treatment is not appropriately randomised to patients, for instance in clinical trials with invalid blinding schemes. This may be due to a faulty protocol or it may be that doctors involved in the trial are aware of the health status of the patients. In contrast to Example 1, it is the doctor in charge of administering treatment who does the “selecting”: if he believes that the drug works, he will tend to give it to patients he thinks will benefit the most. The doctor’s “hunch” will thus be a confounder for the effect of the drug, as it will both determine treatment assignment and be predictive of health outcomes. □

Although the situations described above are formally analogous, they differ in focus. In Example 1 the quantity of interest is the effect of participation for those who chose to participate. This is termed the *effect of treatment on the treated* (ETT). This is also sometimes referred to as the average treatment effect on the treated (ATET) Hotz et al. (2006). In contrast, in Example 2 the quantity of real interest is the *average causal effect* (ACE), as this is what is required for FDA drug approval, for example. The average treatment effect can not usually be identified from confounded observational data unless strong additional assumptions are made. However, with weaker assumptions it may still be possible to identify ETT. This may not be exactly what is really wanted, but can provide some useful information on treatment effects.

Potential responses Current statistical approaches to defining and estimating ETT are almost exclusively based on the *potential response* (PR) framework (Rubin, 1974, 1978). Thus Heckman and Robb (1985) introduced

ETT in the following terms:¹

$$\text{ETT} := \text{E}(Y_1 - Y_0 \mid T = 1). \quad (1)$$

Here T is the treatment variable, with value 1 for a subject receiving active treatment and 0 for a control; while Y_1 and Y_0 are the putative “potential responses” (Rubin, 1978) of a subject to each of these treatments. By definition, it is possible to observe at most one of the two potential responses for any given subject—the other then becoming *counterfactual*. Inference about counterfactuals is sensitive to arbitrary and necessarily untestable assumptions (Dawid, 2000).

Equation (1) can also be expressed as

$$\text{ETT} = \text{E}(Y_1 \mid T = 1) - \text{E}(Y_0 \mid T = 1). \quad (2)$$

It appears *prima facie* that, in order for the expectation in (1) to be meaningful, we must have a joint probability distribution for (Y_1, Y_0, T) —or at the very least, using (2), a conditional distribution for $(Y_1 \mid T = 1)$ and one for $(Y_0 \mid T = 1)$. However since we can never observe Y_0 when $T = 1$, learning the latter distribution from data—and, thus, learning ETT—appears, on the face of it, problematic.

Another approach The above formulation employs counterfactual logic and assumptions. We consider that the current evaluation literature is unnecessarily complicated by the many typically untestable assumptions² needed to use counterfactuals.

Our principal aim in this paper is to demonstrate how a different formalism, eschewing counterfactuals, can helpfully be used to interpret and identify ETT. Our approach is grounded on the *decision theoretic* (DT) framework for causal inference introduced by Dawid (2002, 2007). This supplies a formal language by means of which causal questions can be rigorously posed and analysed, using clear and meaningful assumptions; moreover, fewer such assumptions are typically required than for other approaches such as PR. In

¹Their general definition of ETT allows for further conditioning on a set of observed covariates X , as well as on T . For simplicity we shall omit X wherever this does not affect the thrust of our argument.

²Some of these are *ignorability* and the *stable unit-treatment value assumption* (Rubin, 1986). The former requires that the counterfactual outcomes be independent of the treatment received, the latter requires that the potential response to a treatment of one individual be well-defined, independently of the treatments assigned to other individuals. Another assumption usually invoked in counterfactual theory is *consistency* (Robins, 1986), which requires that the realised response, when treatment t is actually applied, be the same as the corresponding potential response.

the DT framework, causal assumptions are expressed in terms of *conditional independence* statements, that can, in principle if not always in practice, be tested, since all quantities involved are jointly observable. Thus the DT framework provides a more concise, economical and justifiable approach to inference on treatment effects.

We shall present two alternative descriptions of ETT in decision-theoretic terms, and show that they are equivalent. In particular, we show that, contrary to first impressions, ETT is well-defined, being fully determined by the probabilistic behaviour of observable variables. We further show how the PR framework can be formally subsumed within the DT framework as a special case, and deduce that (again, contrary to first impressions) the PR formulation of ETT is itself well-defined in this sense.

Outline The paper is laid out as follows. Section 2 introduces the basic principles of the DT framework. In § 3 we provide a DT definition of ETT in terms of a “preference variable”, governing treatment choice whenever that can be exercised freely, and show that ETT can be identified from observational and interventional data. In § 4 we develop an alternative DT account based on a “sufficient covariate”, and prove that this leads to a unique definition, also agreeing with the earlier one. Moreover, the traditional PR account can be subsumed as a special case of this treatment, and so must deliver the identical answer. In § 5 we provide a formal DT description of two techniques, *matching* and the use of *instrumental variables*, for identifying ETT from observational data alone. We make some concluding remarks in §§ 6 and 7.

2 Decision-theoretic approach to causal inference

The decision-theoretic approach to causal inference (Dawid, 2002, 2003; Dawid and Didelez, 2005; Dawid, 2007) is grounded in the statistical theory of decision-making under uncertainty (Raiffa, 1968; Smith, 1988). Rather than split the response Y of a subject into several potential responses, we consider a variety of stochastic behaviours for the single variable Y (jointly with other relevant variables), under various different *regimes* that may be operating. Our principal purpose is to identify and compare the distributions of Y for a variety of contemplated *interventional* regimes. However data may only have been collected under some *observational* regime. From this standpoint, the major problems to be addressed are whether, when and how probabilistic

information can usefully be transferred across regimes.

For simplicity we restrict attention here to a comparison of two treatments, “active” and “control”, and three different regimes, one observational and two interventional. These reflect, respectively, the circumstances in which a particular observational study of interest is conducted, and those in which one or other of the treatments is administered to a subject. Within each regime, subjects are regarded as exchangeable. We can consequently regard their values on all relevant variables as being drawn, independently across subjects, from some fixed (though generally unknown) joint distribution—which will however generally differ across regimes. The real-world meaning of these regimes will necessarily be context-specific, and the plausibility of any assumptions that may be made about them must be assessed in relation to those real-world meanings.

It is the interventional regimes that are the objects of principal interest, and about which we should like to learn from data, since these will be of direct relevance for guiding future action or policy choice. Thus the government, in deciding whether or not to introduce a new initiative, would want to assess, and compare, the consequences both of action and of inaction. A patient, faced with the decision as to whether or not to take a treatment, needs to assess what the response might be if he did, or if he did not. In either case, knowledge of the distribution of the response under each proposed intervention is exactly what is required to support rational choice between the options. But these interventional distributions may be difficult to identify directly from data collected under observational conditions. One might, naïvely, regard the observational distribution of the response, among those patients who happened to get the treatment, as directly informative about a new patient’s response, if he were to decide to take the treatment; but this would be valid only if he could consider himself exchangeable (on relevant pre-treatment variables) with that observational group. Likewise, for the control group data to be directly relevant for this patient, he would need to regard himself as exchangeable with the observational control group. However it will be impossible to satisfy both conditions simultaneously if—as is common in observational studies—those two groups are not even exchangeable with each other (and not necessarily appropriate even when they are). Then some more refined analysis, typically requiring extra assumptions to be imposed and justified, becomes essential.

2.1 Formal set-up

Denote the treatment variable by T , taking value 1 for active treatment, and 0 for control treatment. We introduce a further variable F , the *intervention*

variable or *regime indicator*. The possible values for F are \emptyset , 0 and 1, indexing the regimes under consideration. When $F = \emptyset$, this indicates that variables are being generated under observational conditions, whereas $F = t$ ($t = 0, 1$) indicates that they are generated under an intervention that sets $T = t$.

Whereas T and Y are chance variables, F is a *decision* or *parameter* variable, and has no uncertainty associated with it. In particular, this means that all probability statements made must be explicitly or implicitly conditional on F . We denote the distribution [resp., expectation] of the chance variables under regime $F = \tau$ ($\tau = 0, 1, \emptyset$) by $p_\tau(\cdot)$ [resp., $E_\tau(\cdot)$]. We note that, under our interpretation of F , we must have, for $t = 0, 1$:

$$F = t \Rightarrow T = t, \tag{3}$$

so that

$$p_t(T = t) = 1 \quad (t = 0, 1). \tag{4}$$

Average causal effect The *average causal effect* (ACE) (of treatment $T = 1$, relative to treatment $T = 0$) on Y is defined as follows:

$$\text{ACE} = E_1(Y) - E_0(Y). \tag{5}$$

This is a simple comparison of the expected value of Y under intervention to apply treatment 1 with that under intervention to apply treatment 0. When utility is linear in the value of the outcome Y , a rational subject with no additional relevant information would prefer treatment 1 to treatment 0 if and only if $\text{ACE} > 0$.³

No confounding In some cases (e.g., randomised trials) we might be prepared to assume that the following conditional independence relation (Dawid, 1979, 2000, 2002) holds:

$$Y \perp\!\!\!\perp F \mid T. \tag{6}$$

This says that, for $t = 0, 1$, $p_\emptyset(y \mid T = t) = p_t(y \mid T = t)$ ($= p_t(y)$, by (3)); *i.e.* given either treatment, the distribution of the response is assumed the same in both the observational regime and the relevant interventional regime. This is the case of “no confounding”, when, for the purpose of estimating the distributions of Y given T , we can treat the observational regime exactly as if

³We could, without adding any real complication, replace Y in (5) by some function of Y , *e.g.* a non-linear measure of the utility of outcome Y . Still more generally, we could compare some other chosen feature, *e.g.* variance, of the distributions of Y under the two experimental regimes $F = 1$ and $F = 0$.

it had been interventional. When (6) holds, $ACE = E_{\theta}(Y | T = 1) - E_{\theta}(Y | T = 0)$, and so can be identified directly from observational data.

The conditional independence assumption (6) can be represented graphically by means of the *influence diagram* (Dawid, 2002) of Figure 1. This is a decision-theoretic version of a directed acyclic graph (DAG), with chance variables represented by round nodes, and decision variables by square nodes. Associated with the arrow from F to T is a specification of the distribution of T in each regime specified by F (in fact degenerate for $F = 0$ or 1 , though non-degenerate for $F = \emptyset$). Associated with that from T to Y is a specification of the conditional distribution of the response Y , given that treatment T has been administered. The *absence* of any arrow from F to Y encodes assumption (6): that this conditional distribution does not further depend on which regime is in operation. (Note however that the property (4) is not encoded in the graph, and has to be introduced explicitly when needed.) Finally we remark that, since F is a decision variable, no probability distribution is associated with it.



Figure 1: Influence diagram representing the conditional independence assumption $Y \perp\!\!\!\perp F | T$

In most contexts (6) can *not* reasonably be assumed (the case of *confounding*). It is such cases that form the focus of this paper.

3 Decision-theoretic formulation of ETT.

I. Preference variable

3.1 Treatment allocation and treatment application

When we compare the responses in two or more treatment group in an observational setting, what we actually see is a combination of two quite distinct effects:

Treatment effect The specific power of the treatment to make a difference to the outcome of interest

Selection effect The fact that we are not observing random subsets of the population of interest

In particular, even if there were no treatment effect whatsoever, the existence of a differential selection effect would typically lead to systematic differences between the outcomes in the different treatment groups, because we would not be comparing like with like: this is the essence of the problem of confounding.

We will find it helpful to elaborate our description and notation to make the above important distinction explicit. We introduce a *preference variable* D , describing the treatment which an individual would choose/be chosen to take if free to do so. Quite separately we have the *treatment variable* T , which indicates which treatment is actually taken or applied. In the observational setting the preferred treatment will be applied, $T = D$: thus D and T are completely confounded with each other. However, in an interventional setting we could override the initial preference, and so need not have $T = D$.

A similar construction was used by Robins et al. (2006). However their analysis involved an additional latent variable U . In § 4, in an alternative to our approach above, we too introduce such a variable U —but as we never need to consider *both* U and D , our description and analysis are more straightforward.

Consider Example 1. We are interested in the effect of the programme on the income of potential participants, *i.e.*, ETT. Each eligible individual, once made aware of the opportunity, makes a personal choice, D , whether to participate or not. We now consider two scenarios. In the first, participation is voluntary: this is the observational regime, $F = \emptyset$, in which $T = D$. The second is the interventional setting: *e.g.* a controlled trial run by the local government that randomises all eligible adults to participate or not in the programme, irrespective of their personal preferences. This gives rise to two interventional regimes, $F = t$ ($t = 0, 1$). These would also be relevant to the considerations a new subject who needs to decide whether or not to participate.

In the context of Example 2, the observational regime, $F = \emptyset$, describes the scenario in which the doctor has a treatment preference D , based perhaps on his hunches about the patient’s likely recovery, and then actually gives the corresponding treatment. The interventional regimes $F = t$ ($t = 0, 1$) refer to the case where the hospital management overrules the doctor’s preference and administers treatment t , or to the treatment decision problem faced by a new patient.

3.2 Assumptions

Certain variables can be classified as “pre-treatment” variables: their values are supposed fully determined before the point at which treatment is actually

applied. In particular this applies to a “covariate”, *i.e.* a permanent pre-existing attribute of a subject. Other variables will be “post-treatment” variables, arising only after the point of treatment application. Although this is not a formal description, in most contexts it will be clear whether a variable belongs to one or other of these groups. Our models will involve only pre-treatment variables, together with a single post-treatment variable, the response Y —and, of course, the treatment variable T itself.⁴

We require the following assumptions:

- i. D is a pre-treatment variable.
- ii. Pre-treatment variables have the same distribution in all regimes, interventional and observational.
- iii. A subject’s response to a treatment does not depend on whether the treatment is self-selected or externally imposed.

Assumption (i) is usually plausible, since it will be reasonable to suppose that, for example:

- An individual knows whether he would wish to participate in a programme, before, and regardless of whether, he is forced to participate or not participate.
- A doctor can decide which treatment he would like to assign to whom, before, and regardless of, whatever the hospital management might decide to do.

At first glance assumption (ii) appears tantamount to assuming “no backwards causation”, which appears entirely reasonable. However, it can fail if, for example, the observational study was done in one population but the proposed interventions relate to a new population, or for a subject who can not be treated as exchangeable (on pre-treatment variables) with those in the study. Thus assumption (ii) also incorporates a requirement of *external validity*.

Assumption (iii) may well be unreasonable in some (especially economic or sociological) contexts, since a subject might behave differently if forced to take a treatment from how he would if he himself had chosen to do so. However little progress can be made without such an assumption.

⁴Thus we do not here consider situations such as that treated by Robins (1989), involving sequential decisions based on accruing time-varying information.

3.3 Regimes

Suppose we have fully specified the preference variable D , and the joint distribution of all relevant observables in the two interventional regimes $F = t$ ($t = 0, 1$). By assumption (ii), these must agree when restricted to pre-treatment variables—including, by assumption (i), D . We can now explicitly construct the joint observational distribution, $F = \emptyset$, of all variables as follows. We first generate all relevant pre-treatment variables, including D , from their joint distribution (which, by (ii), is the same in both interventional regimes); and then use the realised value of D to determine which treatment T to give.⁵ Finally, if, *e.g.*, $D = 1$, we assign to Y the distribution it would have in the active treatment regime, $F = 1$, *conditional on* $D = 1$ (here we use assumption (iii)).

Note that this process can not be reversed: knowing only the observational distribution, in which necessarily $T = D$, we can not in general identify *e.g.* the distribution of Y given $D = 0, T = 1$, which is a necessary ingredient of the interventional regime $F = 1$.

It is easy to see that D has the following properties:

$$D \perp\!\!\!\perp F \tag{7}$$

$$Y \perp\!\!\!\perp F \mid (D, T). \tag{8}$$

Here (7) says that D has the same distribution in all regimes, be they interventional or observational; while (8) says that the distribution of Y given D and T is the same in all regimes, whenever it is meaningfully defined: because of the deterministic dependence of T on F and D , this means that, for $t = 0, 1$, the distribution of Y given $D = t$ is the same in the observational regime $F = \emptyset$ and the interventional regime $F = t$.

The *effect of treatment on the treated* is now defined as:

$$\text{ETT} := E_1(Y \mid D = 1) - E_0(Y \mid D = 1). \tag{9}$$

This is essentially ACE, as defined in (5), but calculated for a specific sub-population of patients: those having $D = 1$, *i.e.*, those who would choose/be chosen to receive treatment 1—whether or not they actually receive it.

We remark that (9) displays a clear separation of “treatment effect” from “differential selection effect”. The former is effected by the comparison of expected responses under the two interventional regimes $F = 1$ and $F = 0$; the latter is excluded because we only compare interventional regimes, and condition on the identical property (namely preference for active treatment, $D = 1$) in both.

⁵Note that T is functionally determined by D and F : $T = t$ when $F = t$ ($t = 0, 1$), and $T = D$ when $F = \emptyset$.

3.4 ETT is identifiable

In a randomised controlled trial, if we could record the preference variable D for all subjects, we could identify ETT straightforwardly. In practice, however, we will not usually be able to observe D in interventional regimes—although we can do so indirectly in observational regimes, since then we know $D = T$, and T is observed. It thus appears *a priori* that it would be impossible to identify the term $E_0(Y \mid D = 1)$ from available data, and thus impossible to identify ETT.

The following analysis shows that, contrary to this initial appearance, ETT *can* be identified—so long as we can gather data under both observational and (some) interventional circumstances. (We must also suppose $p_\emptyset(T = 1) > 0$.)

The first term in (9), $E_1(Y \mid D = 1)$, presents no difficulty. We have:

$$\begin{aligned} E_1(Y \mid D = 1) &= E_1(Y \mid D = 1, T = 1) \\ &= E_\emptyset(Y \mid D = 1, T = 1) \\ &= E_\emptyset(Y \mid T = 1) \end{aligned} \tag{10}$$

where the first equality holds because $F = 1 \Rightarrow T = 1$, the second from (8), and the third because $T = 1 \Rightarrow D = 1$ under $F = \emptyset$. Thus $E_1(Y \mid D = 1)$ is directly identifiable from observational data on (T, Y) .

To get a handle on the problematic second term of (9), $E_0(Y \mid D = 1)$, we argue as follows. We have

$$E_0(Y) = E_0(Y \mid D = 0) \times p_0(D = 0) + E_0(Y \mid D = 1) \times p_0(D = 1). \tag{11}$$

From data gathered under “control” conditions, $F = 0$, we can identify the left-hand side of (11), $E_0(Y)$.

From (8), $p_0(D = 0) = p_\emptyset(D = 0)$, and this in turn is $p_\emptyset(T = 0)$, since $D = T$ in the observational regime $F = \emptyset$. Likewise, $p_0(D = 1) = p_\emptyset(T = 1)$. So these terms can be identified from observational data.

Suppose for the moment (an extreme special case) that $p_\emptyset(T = 1) = 1$, whence $p_\emptyset(T = 0) = 0$. Then from (11) we deduce $E_0(Y \mid D = 1) = E_0(Y)$. Also, (10) becomes $E_\emptyset(Y)$. We thus have, from (9),

$$\text{ETT} = E_\emptyset(Y) - E_0(Y). \tag{12}$$

In this case the observational group behaves just like an interventional treatment group, $E_\emptyset(Y) = E_1(Y)$, and $\text{ETT} = \text{ACE}$.

Otherwise, we have, in parallel fashion to (10), $E_0(Y \mid D = 0) = E_\emptyset(Y \mid T = 0)$, which can be identified from observational data on (T, Y) . The

remaining term in equation (11), $E_0(Y \mid D = 1)$, can thus be solved for. Since we now have both $E_1(Y \mid D = 1)$ from (10) and $E_0(Y \mid D = 1)$ from (11), we can obtain ETT from (9). Doing the algebra, we obtain:

$$\text{ETT} = \frac{E_\emptyset(Y) - E_0(Y)}{p_\emptyset(T = 1)} \quad (13)$$

—a general form that also includes the special case (12).

It follows that ETT is fully identified by the distributions of the observables (T, Y) in the various regimes: indeed, we can identify ETT so long as, in addition to observational data on Y and T , we have data on Y from experimental subjects under control conditions.

Although based on different assumptions, formula (13) is essentially the same as formula (8.20) in Pearl (2000). In § 4.2 we shall see why this must be.

4 Decision-theoretic formulation of ETT.

II. Unobserved confounder

The above development of ETT relies on the existence and meaningfulness of the preference variable D in all regimes, observational and interventional. While this may be a reasonable assumption in *e.g.* economic contexts, where agents may be supposed to form preferences in accordance with rational principles such as maximisation of expected utility, in other contexts it may seem somewhat far-fetched.

We now present an alternative, more general, construction—which, as we shall see, is fully consistent with that described above based on the preference variable. We consider a (typically multivariate, typically unobserved) covariate U (*i.e.*, a permanent attribute of a subject) that can be considered as determining treatment choice—typically only probabilistically—in the observational regime. For Example 1, U might comprise the personal characteristics of the individuals, their motivation, natural talent, confidence, *etc.* For Example 2, U could represent the attributes of the patient that determine the doctor’s hunches as to who will benefit more from the treatment. Such an unobserved variable U , associated with treatment in the observational regime, will be a *confounder* if it is also predictive of outcome.

In contrast to our analysis in § 3.1, we do not now directly construct the observational regime $F = \emptyset$ from the interventional regimes; rather, we regard it as having an entirely independent existence. Then to make progress we must make (and justify!) assumptions relating this to the interventional

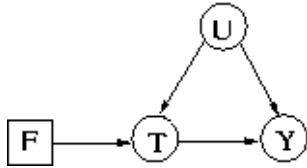


Figure 2: A sufficient covariate U

regimes. Our fundamental requirement is that U be a *sufficient covariate* (Dawid, 2002): that is, for $F \in \{0, 1, \emptyset\}$:

$$U \perp\!\!\!\perp F \tag{14}$$

$$Y \perp\!\!\!\perp F \mid (U, T). \tag{15}$$

Here (14) requires that U (being a pre-treatment variable) have the same distribution in all regimes; while (15) requires that, *if only we knew, and conditioned on*, U , the distribution of the response to an applied treatment would be the same, no matter whether that treatment had been applied under interventional or observational conditions.

The conditional independence relations (14) and (15) can be represented graphically by means of the influence diagram of Figure 2. Note that the arrows in Figure 2 represent stochastic dependence: in particular, T and U need not fully determine Y , but merely modify its distribution. This probabilistic interpretation of a “causal model” may be contrasted with that of (Pearl, 2000§7.1), which would treat U as an undefined exogenous (“error”) variable, and Y as functionally determined by T and U .⁶ Our stochastic model is more general, and appropriate to our intended interpretation of U as a pre-existing real-world attribute of a subject, that could, in principle at least, be identified and measured. It also explicitly allows treatment choice, even in the observational regime, to incorporate an element of randomisation. However, so far as the mathematics is concerned, Pearl’s functional interpretation can be treated as a special case of our model.

Specific causal effect We now introduce the *specific causal effect* of treatment, relative to the specified sufficient covariate U . This is a function of U , defined by

$$\text{SCE}_U := E_1(Y \mid U) - E_0(Y \mid U), \tag{16}$$

i.e.

$$\text{SCE}_U(u) := E_1(Y \mid U = u) - E_0(Y \mid U = u).$$

⁶There appears to be a common perception that graphical models are somehow tied to such a deterministic interpretation. In fact the opposite is true: they are fundamentally tools for manipulating probability distributions, not functional relationships.

That is, $\text{SCE}_U(u)$ is the average causal effect in the subpopulation of individuals having the specified value u for U .

By (15) we can also express

$$\text{SCE}_U = \mathbb{E}_\emptyset(Y \mid T = 1, U) - \mathbb{E}_\emptyset(Y \mid T = 0, U). \quad (17)$$

Thus SCE_U could be identified from observational data if U were to be observed in addition to T and Y .⁷ However typically this will not be the case.

We remark that, by (14), for $t = 0, 1$,

$$\begin{aligned} \mathbb{E}_\emptyset\{\mathbb{E}_t(Y \mid U)\} &= \mathbb{E}_t\{\mathbb{E}_t(Y \mid U)\} \\ &= \mathbb{E}_t(Y), \end{aligned}$$

whence

$$\begin{aligned} \mathbb{E}_\emptyset(\text{SCE}_U) &= \mathbb{E}_1(Y) - \mathbb{E}_0(Y) \\ &= \text{ACE}. \end{aligned} \quad (18)$$

In particular, $\mathbb{E}_\emptyset(\text{SCE}_U)$ can not depend on the choice of sufficient covariate U . Formulas (17) and (18) enable us to identify ACE from an observational study whenever we can measure some sufficient covariate.⁸

4.1 Definition and uniqueness of ETT

After the above preliminaries, we are ready to define the *effect of treatment on the treated* in this setting:

$$\text{ETT}_U := \mathbb{E}_\emptyset\{\text{SCE}_U \mid T = 1\}. \quad (19)$$

That is, ETT_U is the average, in the observational regime, of the specific causal effect (relative to U), for those individuals who in fact receive the active treatment $T = 1$.⁹ This could be identified if we had observational data on all three variables (T, U, Y) . But in general U will not be observable, in any regime: a seemingly fatal handicap to identifying ETT_U .

⁷We also need that, for each value of U , both values of T are observed in the data—which in particular would disallow the choice $U = D$.

⁸Of course, if we could also measure U on a new individual, it would be SCE_U , rather than ACE, that would be relevant for his decision problem.

⁹For this to be meaningful we need to assume $p_\emptyset(T = 1) > 0$. Note also that (16) defines SCE_U only up to a set of probability 0 under the distribution (common to all regimes considered) of U ; but since such a set also has probability 0 in the observational regime conditional on $T = 1$, ETT_U is well-defined.

When a sufficient covariate exists, it need not be unique. The above definition of ETT_U appears, *prima facie*, to depend on the specific choice of sufficient covariate U , and its probabilistic relationship with the observables (T, Y) . However, Theorem 4.1 below (a generalization of the argument of § 3.4) will show that this is not in fact the case: ETT_U does not depend on the choice of U , but only on the joint distributions of the observables in the various regimes. In particular, it can be identified from such data *even when we do not specify, or observe, any sufficient covariate*.

Theorem 4.1 *Suppose $p_\emptyset(T = 1) > 0$. Then, for any sufficient covariate U ,*

$$\text{ETT}_U = \frac{\mathbb{E}_\emptyset(Y) - \mathbb{E}_\emptyset(Y)}{p_\emptyset(T = 1)}. \quad (20)$$

Proof. Suppose first $p_\emptyset(T = 0) > 0$. For $t = 0, 1$, define

$$k(t) := \mathbb{E}_\emptyset\{\mathbb{E}_0(Y | U) | T = t\} \quad (21)$$

$$= \mathbb{E}_\emptyset\{\mathbb{E}_\emptyset(Y | U, T = 0) | T = t\} \quad (22)$$

by (15). In particular, $k(0) = \mathbb{E}_\emptyset(Y | T = 0)$.

By (17) and conditional independence property (15), (19) is equal to

$$\text{ETT}_U = \mathbb{E}_\emptyset(Y | T = 1) - k(1). \quad (23)$$

Also,

$$\begin{aligned} \mathbb{E}_0(Y) &= \mathbb{E}_0\{\mathbb{E}_0(Y | U)\} \\ &= \mathbb{E}_0\{\mathbb{E}_\emptyset(Y | U, T = 0)\} \\ &= \mathbb{E}_\emptyset\{\mathbb{E}_\emptyset(Y | U, T = 0)\} \end{aligned} \quad (24)$$

by (15) and (14). It follows that

$$\begin{aligned} \mathbb{E}_0(Y) &= \mathbb{E}_\emptyset\{k(T)\} \\ &= p_\emptyset(T = 0) k(0) + p_\emptyset(T = 1) k(1). \end{aligned}$$

Hence

$$k(1) = \frac{\mathbb{E}_0(Y) - p_\emptyset(T = 0) \mathbb{E}_\emptyset(Y | T = 0)}{p_\emptyset(T = 1)}. \quad (25)$$

Formula (20) now follows on substituting into (23).

Finally, the special case $p_\emptyset(T = 0) = 0$ can be handled by a similar (and simpler) argument. \square

In the light of the above result, we no longer need to specify which sufficient covariate U is used to define the effect of treatment on the treated; consequently we can just use the notation ETT .

Comments:

- i. Our analysis in § 3.1 in terms of the preference variable D can be treated as a special case of that above, on identifying U with D .
- ii. Suppose that in fact there is no confounding. In that case $E_\emptyset(Y) = E_0(Y) \times p_\emptyset(T = 0) + E_1(Y) \times p_\emptyset(T = 1)$, and formula (20) reduces to $ETT = ACE$.
- iii. It is surprising, though of no ultimate significance, that to identify ETT we do not need data on subjects receiving the treatment by intervention.
- iv. The fact that ETT is well-defined in terms of observable distributions in the various regimes does *not* mean that, in Example 1, it would be the same for two communities with different population distributions and attitudes—since the relevant observational regimes would be different. Similarly in Example 2, different distributions of patients, or different behaviour of the doctors, would typically yield different values for ETT. It is thus a matter for careful consideration whether, and under what circumstances, ETT will be informative about subjects who can not be regarded as exchangeable with those in the study population.
- v. Although the value of ETT does not depend on which sufficient covariate U is being considered, its definition and interpretation require that some such variable U should exist. In some contexts we might not be willing to accept this assumption. Then—notwithstanding the fact that we could still calculate the right-hand side of (20) from knowledge of the distributions of observables in the observational and control regimes—we perhaps should not attempt to interpret this as “the effect of treatment on the treated”.

4.2 Application to potential response framework

In the potential response framework we conceive of the existence of the pair of potential responses (Y_0, Y_1) . These are implicitly supposed to have the same values, and the same joint distribution, no matter what regime operates. That is:

$$(Y_0, Y_1) \perp\!\!\!\perp F. \tag{26}$$

It is also assumed that, no matter what regime operates, the actual response Y is fully determined by the pair (Y_0, Y_1) and the treatment T applied: $Y = Y_T$. This functional dependence implies, in particular:

$$Y \perp\!\!\!\perp F \mid (Y_0, Y_1, T). \quad (27)$$

Comparing (26) and (27) with (14) and (15), we see that we can formally treat $U^* = (Y_0, Y_1)$ as a sufficient covariate. Since $E_t(Y \mid U^*) = Y_t$, the associated specific causal effect is $SCE^* = Y_1 - Y_0$, and hence the associated definition of ETT is $ETT^* = E_\theta(Y_1 - Y_0 \mid T = 1)$. This recovers the “traditional” definition (1) of ETT in the potential response framework.

Now Theorem 4.1 shows that:

- i. The PR definition can be expressed as in (20).
- ii. It agrees with the definition of ETT given in §3.1, as well as with any of the variations, as in (19), in terms of an arbitrary sufficient covariate U .

5 Identifying ETT from observational data

It follows from (20) that we could readily identify ETT if, in addition to having data on T and Y from an observational study, we had data on Y from a group of subjects (randomly selected from the same population) who were made to take the control treatment. However, in many real problems we will not have access to such a control group. Then to make further progress we will need to make, and justify, stronger assumptions, and develop appropriate techniques. Here we consider the use of *instrumental variable* (IV) methods (Heckman and Navarro-Lozano, 2004; Heckman, 2005; Didelez and Sheehan, 2007).

5.1 Identification using an instrumental variable

Consider Example 2 with a twist. The doctor visits a group of patients and, based on his initial diagnosis, he prescribes some of them the new drug. After the doctor has completed this process, these patients are given a preliminary allergy test, and a subset group of them is identified as allergic to some of the components of the drug. As a consequence, the doctor’s prescription is overruled for these patients and they are not administered the drug.

We assume:

- i. The presence of the allergy is independent of the doctor’s treatment preference.

- ii. Conditional on the doctor’s treatment preference, response to control treatment (in this case, not taking the drug) is independent of the presence of allergy.

Property (i) will be plausible when the doctor can not tell which patients are allergic just by talking to them; it also requires that the patient’s medical records do not contain information on allergies to the drug ingredients. This might be the case for instance, when the drug is new, or the ingredients it contains are not in commonly available medication and thus allergy has not been reported. Property (ii) will be plausible if the physiological systems responsible for the allergy and the disease under study are unconnected. Thus, for patients who do not receive the treatment (whether due to the doctor’s treatment preference, or because they have the allergy), response will not be systematically different amongst those who have the allergy. When (i) and (ii) hold, the allergy status of a patient is an *instrumental variable*.

In this case we can use the allergic patients, from whom treatment is withheld, as a proxy for an experimental control group, and so estimate $E_0(Y)$ from these; the remaining patients form, essentially, an observational study group, allowing us to estimate $E_0(Y)$ and $p_0(T = 1)$. Hence we can identify ETT from (20).

5.1.1 Formal construction

Informal arguments such as the above can be valuable and revealing, but it is also necessary to have a rigorous formal system to express and derive such results. To demonstrate how this is provided by our decision-theoretic framework, we present the formal construction below.

We frame our argument in the set-up of §3 (a similar argument could be applied for that of §4). We have variables F, D, T, Y as before—except now we only need to consider $F \in \{0, \emptyset\}$. In addition we have a binary pre-treatment variable A (indicating presence of allergy) such that $p(A = 0) > 0$, and *in the observational regime* $F = \emptyset$, whatever be the value of U :

$$A = 0 \Rightarrow T = 0. \tag{28}$$

In particular, when $A = 0$ this overrides the original treatment preference D . Clearly D , being a pre-treatment variable, continues to satisfy (7). However our original argument for (8) no longer stands. We replace it by

$$Y \perp\!\!\!\perp F \mid (D, T = 0), \tag{29}$$

which says that, conditional always on treatment preference, the observational distribution of response among those who receive the control treat-

ment is the same as its distribution in the control interventional regime. In particular this will hold if D is a sufficient covariate.

We further assume that *in the observational regime* $F = \emptyset$:

$$A \perp\!\!\!\perp D \tag{30}$$

$$Y \perp\!\!\!\perp A \mid (D, T = 0). \tag{31}$$

Here (30) and (31) formalise (i) and (ii) above, respectively. Property (30) must in fact hold in all regimes, since A and D are pre-treatment variables. Condition (31) could have been imposed for active as well as control treatment, but this turns out not to be required for our analysis below.

Note that (28), (30) and (31) are analogous to (3) (for $t = 0$), (7) and (29), respectively, but with A replacing F .

Theorem 5.1 *Under the above conditions, $E_\emptyset(Y \mid A = 1) = E_0(Y)$.*

Proof.

$$\begin{aligned} E_\emptyset(Y \mid A = 0) &= E_\emptyset\{E_\emptyset(Y \mid D, A = 0) \mid A = 0\} \\ &= E_\emptyset\{E_\emptyset(Y \mid D, A = 0)\} \\ &\quad \text{by (30)} \\ &= E_\emptyset\{E_\emptyset(Y \mid D, A = 0, T = 0)\} \\ &\quad \text{by (28)} \\ &= E_0\{E_\emptyset(Y \mid D, A = 0, T = 0)\} \\ &\quad \text{by (7)} \\ &= E_0\{E_\emptyset(Y \mid D, T = 0)\} \\ &\quad \text{by (31)} \\ &= E_0\{E_0(Y \mid D, T = 0)\} \\ &\quad \text{by (29)} \\ &= E_0\{E_0(Y \mid D)\} \\ &\quad \text{by (3)} \\ &= E_0(Y). \end{aligned}$$

□

Using (13), the above result now enables us to identify ETT from data collected under purely observational conditions. The proof extends trivially to whole distribution of Y , not merely its expectation, allowing us to consider alternative loss functions.

5.1.2 IV identification in practice

For the above identification of ETT to work, we only need use an IV to create a proxy for the response of an experimental control group. We are thus more likely to be able to find an appropriate instrument than when IV methods are used to identify ACE, requiring both experimental treatment and control proxies.

A disadvantage is that (31) cannot be empirically tested if, as would be common, we can not observe D when $A = 0$. We will often need to rely on bold and debatable arguments as to the suitability of a purported IV.

For example, Denmark has recently outlawed the use of trans-fats (trans fatty acids) in packaged foods (Stender et al., 2006). Thus, if we wanted to investigate the effect of trans-fats on some health outcomes in Nordic countries, where diets might plausibly be assumed similar, we might treat “being danish” as an IV, so using a random sample of the Danish population as a proxy for an experimental control group, and a random sample of the population of other Nordic countries as the non-experimental treatment group. The assumption of similar diets and lifestyles is debatable, and trans-fats are so widespread in packaged foods that it might be difficult to find a large enough sample of non-experimental untreated. However, it is a plausible approach that might provide valuable information.

5.2 Matching and control functions

Two other methods commonly used to identify the ETT from observational data are *matching* and *control functions* (Heckman and Navarro-Lozano, 2004; Heckman and Vytlačil, 2005; Rubin, 2006), developed for use in the context of labour economics.

Matching essentially defines the problem away by assuming we have an observable sufficient covariate, so allowing identification of ETT by (19), or indeed of ACE by (18).

Control functions seek to relate the observed variables to unobserved variables *via* deterministic functions—in particular, in an econometric setting *personal utility functions*, which are used to measure the likelihood of self-selection, are of crucial importance. Although the method of control functions can be given a DT formulation, we do not consider it further here, since we seek to avoid strong assumptions about unobservable deterministic relations. We do however recognise that in particular contexts, such as the economic problems for which they were introduced, such strong assumptions may be acceptable.

6 Discussion

We have described two related ways in which the concept of “effect of treatment on the treated” can be given a meaningful decision-theoretic interpretation. The first operates by distinguishing between “selection for treatment” and “receipt of treatment”. The second makes use of the existence of an unobserved “sufficient covariate” U . We have shown that the latter approach yields a unique value for ETT, no matter what choice may be made for U , and that this agrees with the value delivered by the former approach. There is also a formal connexion with the approach based on potential responses, ensuring agreement with that too. We have further shown that ETT can be identified so long as we have data both from the observational regime and from an experimental control group.

Identifying ETT in our DT framework relies only on the standard probabilistic machinery once the assumptions (i)-(iii) in § 3.2 are deemed to hold. We do not need to impose additional assumptions, required in the potential response framework to construct counterfactuals, such as consistency (Robins, 1986) or stable unit-treatment value assumption, (Rubin, 1974). The DT approach should be more acceptable to those who appreciate the overall stochastic emphasis of the enterprise of statistical science, since it does not demand a deterministic understanding of causality such as advocated by *e.g.* Heckman (2005).

Whether our assumptions (i)-(iii) in § 3.2, or alternatively (14)-(15) in § 4, are appropriate will depend on the context under consideration, and must be evaluated in the light of specific information. But since these assumptions relate only to the actual world, not to counterfactual parallel worlds, they are relatively straightforward to think about.

7 Postscript

In private correspondence, Judea Pearl has expressed dissatisfaction with our reformulation of ETT as, essentially, “the effect of treatment on the treatable”. He insists that his own treatment (see Pearl (2000), § 8.2.5) addresses what he takes to be the real question of interest: “What would have happened to those who actually got treated, if they had not got treated?”.

Now whether this question is answerable depends on the information at hand. We already know, and should therefore condition on, the actual outcomes of those who were treated. But neither our analysis nor Pearl’s can handle that conditioning: indeed there is no identifiable answer to this question that does not depend on additional untestable assumptions about the

relationship between the potential responses for treatment and for control (Dawid, 2000).

But suppose, instead, we agree to blind ourselves to these data on outcomes; or alternatively we have based our estimates on a random sample of those who actually got treated, and wish to apply Pearl's question to the remainder, on whom we do not have outcome data. The target individuals have indeed taken the treatment, and so it might indeed be classed as counterfactual to consider what might have happened to them if they had not. Nevertheless, in the absence of any outcome or other post-treatment information, from an epistemological viewpoint we have exactly the same data from them (*i.e.*, none at all, in our simple version, or pre-existing covariate data in more elaborate versions) as we would have from a new patient who has been fingered for treatment but is yet to be treated. The nice philosophical distinction between the two cases is thus of no practical concern, and the mathematics of our approach is equally relevant to both (indeed, assumptions (ii) of § 3.2, or (14) of § 4, are all the more plausible in this new story).

References

- Dawid, A. P. (1979). Conditional independence in statistical theory (with Discussion). *Journal of the Royal Statistical Society, Series B* 41, 1–31.
- Dawid, A. P. (2000). Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association* 95, 407–448.
- Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review* 70, 161–189. Corrigenda, *ibid.*, 437.
- Dawid, A. P. (2003). Causal inference using influence diagrams: The problem of partial compliance (with Discussion). In P. J. Green, N. L. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, pp. 45–81. Oxford University Press.
- Dawid, A. P. (2007). Fundamentals of statistical causality. Research Report 279, Department of Statistical Science, University College London.
- Dawid, A. P. and V. Didelez (2005). Identifying the consequences of dynamic treatment strategies. Research Report 262, Department of Statistical Science, University College London.

- Didelez, V. and N. A. Sheehan (2007). Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* 16, 309–330.
- Forcina, A. (2006). Causal effects in the presence of non-compliance: A latent variable interpretation (with Discussion). *Metron* 64, 275–302.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* 47, 153–161.
- Heckman, J. (2005). The scientific model of causality. *Sociological Methodology* 35, 1–98.
- Heckman, J. and S. Navarro-Lozano (2004, February). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics* 80, 30–57.
- Heckman, J. and R. Robb (1985). Alternative methods for estimating the impact of interventions. In J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data*, pp. 156–245. New York: Cambridge University Press.
- Heckman, J. and E. Vytlacil (2005). Structural equations treatment effects and econometric policy evaluation. *Econometrica* 73, 669–738.
- Hotz, J. V., G. W. Imbens, and J. A. Klerman (2006). Evaluating the differential effects of alternative welfare-to-work training components: A re-analysis of the California GAIN program. *Journal of Labor Economics* 24, 521–566.
- Okie, S. (2006). Access before approval—a right to take experimental drugs? *New England Journal of Medicine* 355, 437–440.
- Pearl, J. (2000). *Causality*. Cambridge University Press.
- Raiiffa, H. (1968). *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Reading, Massachusetts: Addison-Wesley.
- Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 7, 1393–1512.
- Robins, J. M. (1989). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal

- studies. In L. Sechrest, H. Freeman, and A. Mulley (Eds.), *Health Service Research Methodology: A Focus on AIDS*, pp. 113–159. NCSHR, U.S. Public Health Service.
- Robins, J. M., T. J. VanderWeele, and T. S. Richardson (2006). Comment on “Causal effects in the presence of non compliance: A latent variable interpretation” by Antonio Forcina. *Metron* 64, 288–298.
- Rubin, D. (1986). Comment: Statistics and causal inference. *Journal of the American Statistical Association* 81, 968–970.
- Rubin, D. (2006). *Matched Sampling for Causal Effects*. Harvard University Press.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688–701.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The rôle of randomization. *Annals of Statistics* 6, 34–68.
- Smith, J. Q. (1988). *Decision Analysis: A Bayesian Approach*. London: Chapman and Hall.
- Stender, S., J. Dyerberg, A. Bysted, T. Leth, and A. Astru (2006). A trans world journey. *Atherosclerosis Supplements* 7, 47–52.