Decision-theoretic foundations for statistical causality

A. Philip Dawid*

April 28, 2020

Abstract

We develop a mathematical and interpretative foundation for the enterprise of decision-theoretic statistical causality (DT), which is a straightforward way of representing and addressing causal questions. DT reframes causal inference as "assisted decision-making", and aims to understand when, and how, I can make use of external data, typically observational, to help me solve a decision problem by taking advantage of assumed relationships between the data and my problem.

The relationships embodied in any representation of a causal problem require deeper justification, which is necessarily context-dependent. Here we clarify the considerations needed to support applications of the DT methodology. Exchangeability considerations are used to structure the required relationships, and a distinction drawn between intention to treat and intervention to treat forms the basis for the enabling condition of "ignorability".

We also show how the DT perspective unifies and sheds light on other popular formalisations of statistical causality, including potential responses and directed acyclic graphs.

Key words: directed acyclic graph, exchangeability, extended conditional independence, ignorability, potential outcome, single world intervention graph

1 Introduction

The decision-theoretic (DT) approach to statistical causality has been described and developed in a series of papers (Dawid 2000; Dawid 2002; Dawid 2003; Didelez *et al.* 2006; Dawid 2007a; Geneletti 2007; Dawid and Didelez 2008; Dawid and Didelez 2010; Guo and Dawid 2010; Geneletti and Dawid 2011; Dawid 2012; Berzuini *et al.* 2012b; Dawid and Constantinou 2014; Guo *et al.* 2016); for general overview see Dawid (2007b); Dawid (2015). It has been shown to be a more straightforward approach, both philosophically and for use in applications, than other popular frameworks for statistical causality based *e.g.* on potential responses or directed acyclic graphs.

From the standpoint of DT, "causal inference" is something of a misnomer for the great preponderance of the methodological and applied contributions that normally go by this description. A better characterisation of the field would be "assisted decision making". Thus the DT approach focuses on how we might make use of external—typically observational—data to help inform a decision-maker how best to act; it aims to characterise conditions allowing this, and to develop ways in which it can be achieved. Work to date has concentrated on the nuts and bolts of showing

---

*University of Cambridge

how the DT approach may be applied to a variety of problems, but has largely avoided any detailed consideration of how the conditions enabling such application might be justified in terms of still more fundamental assumptions. The main purpose of the present paper is to to conduct a careful and rigorous analysis, to serve as a foundational "prequel" to the DT enterprise. We develop, in detail, the basic structures and assumptions that, when appropriate, would justify the use of a DT model in a given context—a step largely taken for granted in earlier work. We emphasise important distinctions, such as that between cause and effect variables, and that between intended and applied treatment, both of which are reflected in the formal language; another important distinction is that between post-treatment and pre-treatment exchangeability. The rigorous development is based on the algebraic theory of extended conditional independence, which admits both stochastic and non-stochastic variables (Dawid 1979a; Dawid 1980; Constantinou and Dawid 2017), and its graphical representation (Dawid 2002).

We also consider the relationships between DT and alternative current formulations of statistical causality, including potential outcomes (Rubin 1974; Rubin 1978), Pearlian DAGs (Pearl 2009), and single world intervention graphs (Richardson and Robins 2013a; Richardson and Robins 2013b). We develop DT analogues of concepts that have been considered fundamental in these alternative approaches, including consistency, ignorability, and the stable unit-treatment value assumption. In view of these connexions, we hope that this foundational analysis of DT causality will also be of interest and value to those who would seek a deeper understanding of their own preferred causal framework, and in particular of the conditions that need to be satisfied to justify their models.

Plan of paper

Section 2 describes, with simple examples, the basics of the DT approach to modelling problems of "statistical causality", noting in particular the usefulness of introducing a non-stochastic variable that allows us to distinguish between the different regimes—observational and interventional—of interest. It shows how assumed relationships between these regimes, intended to support causal inference, may be fruitfully expressed using the language and notation of extended conditional independence, and represented graphically by means of an augmented directed acyclic graph.

In § 3 and § 4 we describe and illustrate the standard approach to modelling a decision problem, as represented by a decision tree. The distinction between cause and effect is reflected by regarding a cause as a non-stochastic decision variable, under the external control of the decision-maker, while an effect is a stochastic variable, that can not be directly controlled in this way. We introduce the concept of the hypothetical distribution for an effect variable, were a certain action to be taken, and point out that all we need, to solve the decision problem, is the collection of all such hypothetical distributions.

Section 5 frames the purpose of "causal inference" as taking advantage of external data to help me solve my decision problem, by allowing me to update my hypothetical distributions appropriately. This is elaborated in § 6, where we relate the external data to my own problem by means of the concept of exchangeability. We distinguish between post-treatment exchangeability, which allows straightforward use of the data, and pre-treatment exchangeability, which can not so use the data without making further assumptions. These assumptions—especially, ignorability—are developed in § 7, in terms of a clear formal distinction between intention to treat and intervention to treat. In § 8 we develop this formalism further, introducing the non-stochastic regime indicator that is central to the DT formulation. Section 9 generalises this by introducing additional covariate information, while § 10 generalises still further to problems represented by a directed acyclic graph.

2

In §11 we highlight similarities and differences between the DT approach to statistical causality and other formalisms, including potential outcomes, Pearlian DAGs, and single-world intervention graphs. These comparisons and contrasts are explored further in §12, by application to a specific problem, and it is shown how the DT approach brings harmony to the babel of different voices. Section 13 rounds off with a general discussion and suggestions for further developments. Some technical proofs are relegated to Appendix A.

2   The DT approach

Here we give a brief overview of the DT perspective on modelling problems of statistical causality.

A fundamental feature of the DT approach is its consideration of the relationships between the various probability distributions that govern different regimes of interest. As a very simple example, suppose that we have a binary treatment variable T, and a response variable Y. We consider three different regimes, indexed by the values of a non-stochastic regime indicator variable F_T:[1]

$F_T = 1$. This is the regime in which the active treatment is administered to the patient

$F_T = 0$. This is the regime in which the control treatment is administered to the patient

$F_T = \emptyset$. This is a regime in which the choice of treatment is left to some uncontrolled external source.

The first two regimes may be described as *interventional*, and the last as *observational*. In each regime there will be a joint distribution for the treatment and response variables, T and Y. The distribution of T will be degenerate under an interventional regime (with $T = 1$ almost surely under $F_T = 1$ and $T = 0$ almost surely under $F_T = 0$); but T will typically have a non-degenerate distribution in the observational regime

It will often be the case that I have access to data collected under the observational regime $F_T = \emptyset$; but for decision-making purposes I am interested in comparing and choosing between two interventions available to me, $F_T = 1$ and $F_T = 0$, for which I do not have direct access to relevant data. I can only use the observational data to address my decision problem if I can make, and justify, appropriate assumptions relating the distributions associated with the different regimes.

The simplest such assumption (which, however, will often not be easy to justify) is that the distribution of Y in the interventional active treatment regime $F_T = 1$ is the same as the conditional distribution of Y, given $T = 1$, in the observational regime $F_T = \emptyset$; and likewise the distribution of Y under regime $F_T = 0$ is the same as the conditional distribution of Y given $T = 0$ in the regime $F_T = \emptyset$. This assumption can be expressed, in the conditional independence notation of Dawid (1979a), as:

$$Y \perp\!\!\!\perp F_T \mid T, \tag{1}$$

(read: "Y is independent of F_T, given T"), which asserts that the conditional distributions of the response Y, given the administered treatment T, does not further depend on F_T (*i.e.*, on whether that treatment arose naturally, in the observational regime, or by an imposed intervention), and so can be chosen to be the same in all three regimes.

Note, importantly, that the conditional independence assertion (1) makes perfect intuitive sense, even though the variable F_T that occurs in it is non-stochastic. The intuitive content of (1) is made

---

[1]The use of explicit intervention variables such as F_T was pioneered by Pearl (1993a); Pearl (1993b), although, for reasons obscure to this author, he seems largely to have abandoned it very quickly.

fully rigorous by the theory of extended conditional independence (ECI) (Dawid 1980; Constantinou and Dawid 2017), which shows that such expressions can, with care, be manipulated in exactly the same way as when all variables are stochastic.

Property (1) can also be expressed graphically, by the augmented DAG (directed acyclic graph) (Dawid 2002) of Figure 1. Again, we can include both stochastic variables (represented by round nodes) and non-stochastic variables (square nodes) in such a graph, which encodes extended conditional independence by means of the d-separation criterion (Geiger *et al.* 1990) or the equivalent moralisation criterion (Lauritzen *et al.* 1990). In Figure 1 it is the absence of an arrow from F_T to Y that encodes property (1).
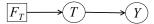

Figure 1: A simple augmented DAG

The identity, expressed by (1), of the conditional distribution of Y given T, across all the regimes described by the values of the regime indicator F_T, can be understood as expressing the *invariance* or *stability* (Bühlmann 2018) of a probabilistic ingredient—the conditional distribution of Y, given T—across the different regimes. This is thus being regarded as a modular component, unchanged wherever it appears in any of the regimes. When it can be justified, the stability property represented by (1) or Figure 1 permits *transfer* (Pearl and Bareinboim 2011) of relevant information between the regimes: we can use the (available, but not directly interesting) observational data to estimate the distributions of response Y given treatment T in regime $F_T = \emptyset$; and then regard these observational conditional distributions as also supplying the desired interventional distributions of Y (of interest, but not directly available) in the hypothetical regimes $F_T = 1$ and $F_T = 0$ relevant to my decision problem.[2] Characterising, justifying, and capitalising on such modularity properties are core features of the DT approach to causality.

A more complex example is given by the DAG of Figure 2, which represents a problem where Z is an instrumental variable for the effect of a binary exposure variable X on an outcome variable Y, in the presence of unobserved "confounding variables" U. Note again the inclusion of the regime indicator F_X, with values 0, 1 and $\emptyset$. As before, $F_X = \emptyset$ labels the observational regime in which data are actually obtained, while $F_X = 1$ [resp., 0] labels the hypothetical regime where we intervene to force X to take the value 1 [resp., 0].

The figure is nothing more nor less than the graphical representation of the following extended conditional independence properties (which it embodies by means of d-separation):

$$(Z, U) \quad \perp\!\!\!\perp \quad F_X \tag{2}$$
$$U \quad \perp\!\!\!\perp \quad Z \quad | \quad F_X \tag{3}$$

---

[2]An important aside on notation and terminology. In the potential outcome (PO) approach, the response Y is artificially split into two, Y_0 and Y_1, it being supposed that Y_t is what is observed in regime F_t—the marginal distribution of Y_t thus being the same as our hypothetical distribution for Y under intervention $F_T = t$. This duplication of the response is entirely unnecessary for our purposes. Moreover, there is a very prevalent misuse of terms such as "counterfactual distribution", or "estimating the counterfactual", notwithstanding that there is nothing counter to any known fact involved in considering these distributions, which are to be applied to a new case. We have termed the interventional distributions of Y *hypothetical*, since they are predicated on a hypothetical intervention on a new case. I have elsewhere (Dawid 2007a) expanded on the importance of distinguishing between hypothetical and counterfactual reasoning, which is jeopardised when we do not also make a clear terminological distinction.
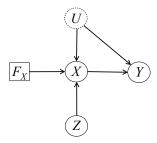
Figure 2: Instrumental variable with regimes

$$Y \quad \perp\!\!\!\perp \quad Z \quad | \ (X, U, F_X) \tag{4}$$
$$Y \quad \perp\!\!\!\perp \quad F_X \ | \ (X, U). \tag{5}$$

In words, (2) asserts that the joint distribution of Z and U is a modular component, the same in all 3 regimes, while (3) further requires that, in this (common) joint distribution, we have independence between U and Z. Next, (4) says that, in any regime, the response Y is independent of the instrument Z, conditionally on exposure X and confounders U (the "exclusion restriction"); while (5) further requires that the conditional distribution for Y, given X and U (which, by (4), is unaffected by further conditioning on Z) be the same in all regimes.

We emphasise that properties (2)–(5) comprise the full extent of the causal assumptions made. In particular—and in contrast to other common interpretations of a "causal graph" (Dawid 2010)— no further causal conclusions should be drawn from the directions of the arrows in Figure 2. In particular, the arrow from Z to X should not be interpreted as implying a causal effect of Z on X: indeed, the figure is fully consistent with alternative causal assumptions, for example that Z and X are merely associated by sharing a common cause (Dawid 2010). In general, the causal content of any augmented DAG is to be understood as fully comprised by the extended conditional independencies that it embodies by d-separation. This gives a precise and clear semantics to our "causal DAGs".

To the extent that the assumptions embodied in Figure 2 imply restrictions on the observational distribution of the data (*i.e.*, properties (3) and (4), considered only under the operation of the observational regime $F_X = \emptyset$), they tally with the standard assumptions made in instrumental variable analysis (Hernán and Robins 2006). However, without the additional stitching together of behaviours under the observational regime and the desired, but unobserved, interventional regimes, it is not possible to use the observational data to make causal inferences. When, and only when, these additional stability assumptions can be made can we justify application of the usual methods of instrumental variable analysis.

In previous work, we have used the above formulation in terms of extended conditional independences, involving both stochastic variables and non-stochastic regime indicators, as the starting point for analysis and discussion of statistical causality, both in general terms and in particular applications. In this work, we aim to dig a little deeper into the foundations, and in particular to understand why, when, and how we might justify the specific extended conditional independence properties previously simply assumed.

3 Causality, agency and decision

There is a very wide variety of philosophical understandings and interpretations of the concept of "causality". Our own approach is closely aligned with the "agency" interpretation (Reichenbach 1956; Price 1991; Hausman 1998; Woodward 2003; Woodward 2016), whereby a "cause" is understood as something that can (at least in principle) be externally manipulated—this notion being an undefined primitive, whose intended meaning is easy enough to comprehend intuitively in spite of being philosophically contentious (Webb 2020). This is not to deny the value of other interpretations of causality, based for example on mechanisms (Salmon 1984; Dowe 2000), simplicity (Janzing and Schölkopf 2010), probabilistic independence (Suppes 1970; Spohn 2001) or invariant processes (Bühlmann 2018), or starting from different primitive notions, such as common cause or direct effect (Spirtes *et al.* 2000), or one variable "listening to" another (Pearl and Mackenzie 2018). However, the present work has the limited aim of explicating the agency-based decision-theoretic approach.

The basic idea is that an agent ("I", say) has free choice among a set of available actions, and that performing an action will, in some sense, tend to bring about some outcome. Indeed, whenever I seriously contemplate performing some action, my purpose is to bring about some desired outcome; and that aim will inform my choice between the different actions that may be available. We may consider my action as a putative "cause" of my outcome. This approach makes a clear distinction between cause and effect: the former is represented as an action, subject to my free choice, while the latter is represented as an outcome variable, over which I have no direct control. Correspondingly, we will need different formal representations for cause and effect variables: only the latter will be treated as stochastic random variables.

Now by my action I generally won't be able to determine the outcome exactly, since it will also be affected by many circumstances beyond my control, which we might ascribe to the vagaries of "Nature". So I will have uncertainty about the eventual outcome that would ensue from my action. We shall take it for granted that it is always appropriate to represent my uncertainty by a probability distribution. Then, for any contemplated but not yet executed action a, there will be a joint probability distribution P_a over all the ensuing variables in the problem[3], representing my current uncertainty (conditioned on whatever knowledge I currently have, prior to choosing my action) about how those variables *might* turn out, *were I to perform* action a. We will term such a distribution P_a *hypothetical*, since it is premised on the *hypothesis* that I perform action a.

There will be a collection $\mathcal{A}$ of actions available to me, and correspondingly an associated collection $\{P_a : a \in \mathcal{A}\}$ of my hypothetical distributions—each contingent on just one of the actions I might take. My task is to rank my preferences among these different hypothetical distributions over future outcomes, and perform that action corresponding to the distribution P_a I like best. I can do this ranking in terms of any feature of the distributions that interests me.

One such way, concordant with Bayesian statistical decision theory (Raiffa and Schlaifer 1961; DeGroot 1970), is to construct a real-valued loss function L, such that $L(y, a)$ measures the dissatisfaction I will suffer if I take action a and the value of some associated outcome variable Y later turns out to be y. This is represented in the decision tree of Figure 3.

The square at node ν_* indicates that it is a decision node, where I can choose my action, a. The round node ν_a indicates the generation of the stochastic outcome variable, Y, whose hypothetical distribution P_a will typically depend on the chosen action a.

---

[3]In full generality, the relevant collection of ensuing variables could itself depend on my action a; purely for simplicity we shall restrict to the case that it does not.
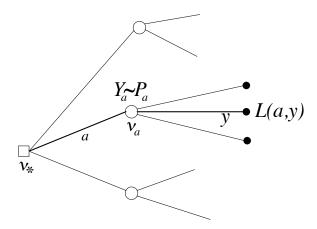
Figure 3: Decision tree

Since, at node ν_a, $Y \sim P_a$, the (negative) value of taking action a, and thus getting to ν_a, is measured by the expected loss $L(a) := E_{Y \sim P_a}\{L(Y, a)\}$. The principles of statistical decision analysis now require that, at the decision node ν_*, I should choose an action a minimising $L(a)$.

Note particularly that, whatever loss function is used, this solution will only require knowledge of the collection $\{P_a\}$ of hypothetical distributions for the outcome variable Y.

There are decision problems where explicit inclusion of the action a as an argument of the loss function is natural. For example, I might have a choice between taking my umbrella ($a = 1$) when I go out, or leaving it at home ($a = 0$). For either action, the relevant binary outcome variable Y indicates whether it rains ($Y = 1$) or not ($Y = 0$). The loss is 1 if I get wet, 0 otherwise, so that $L(0, 0) = L(0, 1) = L(1, 1) = 0$, $L(1, 0) = 1$. In this case, my action presumably has no effect on the outcome Y, so that I might take P_1 and P_0 to be identical; but it enters non-trivially into the loss function. However, it is arguable whether such a problem, where the only effect of my action is on the loss, can properly be described as one of causality. In typical causal applications, the loss function will depend only on the value y of Y, and not further on my action—so that $L(y, a)$ simplifies to $L(y)$. The only thing depending on a will then be my hypothetical distribution P_a for Y, subsequent to ("caused by") my taking action a. Then $L(a) = E_{Y \sim P_a}\{L(Y)\}$, and my choice of action effectively becomes a choice between the different hypothetical distributions P_a for Y associated with my available actions a: I prefer that distribution giving the smallest expectation for $L(Y)$. This specialisation will be assumed throughout this work.

4   A simple causal decision problem

As a simple specific example, we consider the following stylised decision problem.

Example 1 I have a headache and am considering whether or not I should take two aspirin tablets. Will taking the aspirins cause my headache to disappear?

Let the binary decision variable F_X denote whether I take the aspirin ($F_X = 1$) or not ($F_X = 0$), and let Z denote the time it takes for my headache to go away. For convenience only, we focus on $Y := \log Z$, which can take both positive and negative values.

I myself will choose the value of F_X: it is a decision variable, and does not have a probability distribution. Nevertheless, it is still meaningful to consider my conditional distribution, P_x say, for how the eventual response Y would turn out, where I to take decision $F_X = x$ ($x = 0, 1$). For the moment we assume the distributions P_0, P_1 to be known—this will be relaxed in §5. Where we need to be definite, we shall, purely for simplicity, take P_x to have the normal distribution $\mathcal{N}(\mu_x, \sigma^2)$, with probability density function:

$$p_x(y) \equiv p(y \mid F_X = x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp - \frac{(y - \mu_x)^2}{2\sigma^2}, \tag{6}$$

having mean μ_0 or μ_1 according as $x = 0$ or 1, and variance σ^2 in either case.

The distribution P_1 [resp., P_0] expresses my *hypothetical* uncertainty about how Y would turn out, *if* I were to decide to take the aspirin, *i.e.* under $F_X = 1$ [resp., if I were to decide not to take the aspirin, $F_X = 0$]. It can incorporate various sources and types of uncertainty, including stochastic effects of external influences arising or acting between the point of treatment application and the eventual response. My task is to compare the two hypothetical distributions P_1 and P_0, and decide which one I prefer. If I prefer P_1 to P_0, then my decision should be to take the aspirin; otherwise, not. Whatever criterion I use, all I need to put it into effect, and so solve my decision problem, is the pair of hypothetical distributions $\{P_0, P_1\}$ for the outcome Y, under each of my hypothesised actions.

One possible comparison of P_1 and P_0 might be in terms their respective means, μ_1 and μ_0, for Y; the "effect" of taking aspirin, rather than nothing, might then be quantified by means of the change in the expected response, $\delta := \mu_1 - \mu_0$. This is termed the *average causal effect*, ACE (in terms of the outcome variable Y—so more specifically denoted by ACE_Y, if required). Alternatively, we might look at the average causal effect in terms of $Z = e^Y$: $\mathrm{ACE}_Z = \mathrm{E}_{P_1}(Z) - \mathrm{E}_{P_0}(Z) = e^{\sigma^2/2}(e^{\mu_1} - e^{\mu_0})$, or make this comparison as a ratio, $\mathrm{E}_{P_1}(Z)/\mathrm{E}_{P_0}(Z) = e^{\mu_1 - \mu_0}$. Or, we could consider and compare the variance of Z, $\mathrm{var}_x(Z) = e^{2\mu_x}(e^{2\sigma^2} - e^{\sigma^2})$ under P_x ($x = 0, 1$). In full generality, any comparison of an appropriately chosen feature of the two hypothetical distributions, P_0 and P_1, of Y can be regarded as a partial summary of the *causal effect* of taking aspirin (as against taking nothing).

A fully decision-theoretic formulation is represented by the decision tree of Figure 4.
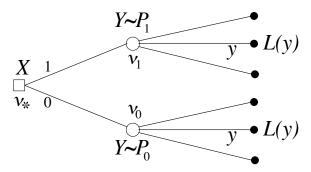

Figure 4: Decision tree

Suppose (for example) that I were to measure the loss that I will suffer if my headache lasts $z = e^y$ minutes by means of the real-valued loss function $L(z) = \log z = y$. If I were to take the aspirin ($F_X = 1$), my expected loss would be $E_{Y \sim P_1}(Y) = \mu_1$; if not ($F_X = 0$), it would be μ_0.

The principles of statistical decision analysis now direct me to choose the action leading to the smaller expected loss. The "effect of taking aspirin" might be measured by the increase in expected loss, which in this case is just ACE_Y; and the correct decision will be to take aspirin when this is negative.

Although there is no uniquely appropriate measure of "the effect of treatment", in the rest of our discussion we shall, purely for simplicity and with no real loss of generality, focus on the difference of the means of the two hypothetical distributions for the outcome variable Y:

$$\mathrm{ACE} = \mathrm{E}_{P_1}(Y) - \mathrm{E}_{P_0}(Y). \tag{7}$$

$\square$

5 Populating the decision tree

The above formulation is fine so long as I know all the ingredients in the decision tree, in particular the two hypothetical distributions P_0 and P_1. Suppose, however, that I am uncertain about the parameters μ_1 and μ_0 of the relevant hypothetical distributions P_1 and P_0 (purely for simplicity we shall continue to regard σ^2 as known). To make explicit the dependence of the hypothetical distributions on the parameters, we now write them as P_{1,μ_1}, P_{0,μ_0}, and denote the associated density functions by $p_1(y \mid \mu_1)$, $p_0(y \mid \mu_0)$.

5.1 No-data decision problem

Being now uncertain about the parameter-pair $\boldsymbol{\mu} = (\mu_1, \mu_0)$, I should assess my personalist prior probability distribution, Π say, for $\boldsymbol{\mu}$ (in the light of whatever information I currently have). Let this have density $\pi(\mu_1, \mu_0)$. To solve my decision problem, I would then substitute, for the unknown hypothetical distribution $P_{1,\mu_1}(y)$, my "prior predictive" hypothetical distribution P_1^* for Y, with density

$$
\begin{aligned}
p_1^*(y) &= \int \int p_1(y \mid \mu_1)\, \pi(\mu_1, \mu_0)\, d\mu_1\, d\mu_0 \\
&= \int p_1(y \mid \mu_1)\, \pi_1(\mu_1)\, d\mu_1
\end{aligned}
$$

where $\pi_1(\mu_1)$ is my marginal prior density for μ_1:

$$\pi_1(\mu_1) = \int \pi(\mu_1, \mu_0)\, d\mu_0.$$

Similarly, I would replace $P_{0,\mu_0}(y)$ by P_0^*, having density $p_0^*(y) = \int p_0(y \mid \mu_0)\, \pi_0(\mu_0)\, d\mu_0$, where $\pi_0(\mu_0) = \int \pi(\mu_1, \mu_0)\, d\mu_1$ is my marginal prior density for μ_0. We remark that, in parallel to the property that, with full information, I only need to specify the two hypothetical distributions P_1 and P_0, when I have only partial information I only need to specify, separately, my marginal uncertainties about the unknown parameters of each of these distributions. In particular, once these margins have been specified, any further dependence structure in my joint personal probability distribution Π for (μ_1, μ_0) is irrelevant to my decision problem.

5.2 Data

When in a state of uncertainty, that uncertainty can often be reduced by gathering data. Bayesian statistical decision theory (Raiffa and Schlaifer 1961) shows that, for any decision problem, the expected reduction in loss by using additional data ("the expected value of sample information") is always non-negative. The effect of obtaining data D is to replace all the distributions entering in §5.1 above by their versions obtained by further conditioning on D.

Suppose then that I wish to reduce my uncertainty about μ_1, the parameter of my hypothetical distribution P_1, by utilising relevant data. What data should I collect, and how should I use them?

What I might, ideally, want to do is gather together a "treatment group" $\mathcal{T}$ of individuals whom I can regard, in an intuitive sense, as similar to myself, with headaches similar to my own. We call such individuals *exchangeable* (both with each other and with me)—this intuitive concept is treated more formally in §6 below. I then give them each two aspirins, and observe their responses (how long until their headaches go away). Conditionally on the parameter μ_1 of $P_1 = P_{1,\mu_1}$, I could reasonably[4] model these responses as being independently and identically distributed, with the same distribution, P_{1,μ_1}, that would describe my own uncertainty about my own outcome, Y, were I, hypothetically, to take the aspirins, and thus put myself into the identical situation as the individuals in my sample. Conditionally on μ_1, I would further regard my own outcome as independent of those in the sample. We shall not here be concerned with issues of sampling variability in finite datasets. So we consider the case that the treatment group $\mathcal{T}$ is very large. Then I can essentially identify μ_1 as the observed sample mean $\widehat{\mu}_1$, and so take my updated P_1 to be $\mathcal{N}(\widehat{\mu}_1, \sigma^2)$.[5] For any non-dogmatic prior, this will be a close approximation to my Bayesian "posterior predictive distribution" for Y, given the data D (conditionally on my taking the aspirins), and also has a clear frequentist justification.

The above was relevant to my hypothetical distribution P_1, were I to take the aspirins. But of course an entirely parallel argument can be applied to estimating P_0, the distribution of my response Y were I not to take the aspirins. I would gather another large group (the "control group", $\mathcal{C}$) of individuals similar to myself, with headaches similar to my own, but this time withhold the aspirins from them. I would then use the empirically estimated distribution of the response in this group as my own distribution P_0.

Let $\mathcal{D} = \mathcal{T} \cup \mathcal{C}$ be the set of "data individuals". Using the responses of $\mathcal{D}$, I have been able to populate my own decision problem with the relevant hypothetical distributions, P_1 and P_0. I can now solve it, and so choose the optimal decision for me.

6 Exchangeability

Here we delve more deeply into the justification for some of the intuitive arguments made above (and below).

In §5.2 above, in the context first of estimating my hypothetical distribution P_1, we talked of constructing, as the treatment group $\mathcal{T}$,

> "a group of individuals whom I can regard, in an intuitive sense, as similar to myself, with headaches similar to my own".

---

[4] See §6 for formal justification.

[5] This is of course specific to our assumed normal model (6), and in any case assumes σ^2 known. For other models we might plug in the maximum likelihood estimate (or any other consistent estimate). Still more generally, we could estimate the distribution P_1 nonparametricly, *e.g.* using the empirical distribution of the sample data.

The identical requirement was imposed on the control group $\mathcal{C}$. The formal definition and theory of *exchangeability* (de Finetti 1937; de Finetti 1975) seeks to put this intuitive conception on a more formal footing.

We consider a collection $\mathcal{I}$ of individuals, on each of which we can measure a number of generic variables. One such is the generic *response variable* Y, having a specific instance, Y_i, for individual i—that is, Y_i denotes the response of individual i. We suppose all individuals considered are included in $\mathcal{I}$. In particular, $\mathcal{T} \subseteq \mathcal{I}$, $\mathcal{C} \subseteq \mathcal{I}$, and I myself am included in $\mathcal{I}$, with label 0, say.

6.1 Post-treatment exchangeability

What we are essentially requiring of $\mathcal{T}$, in the description quoted above, is twofold:

(i). My joint personalist distribution for the responses in the treatment group, *i.e.* the ordered set $(Y_i : i \in \mathcal{T})$, is exchangeable—that is to say, I regard the re-ordered set $(Y_{\rho(i)} : i \in \mathcal{T})$ as having the same joint distribution as $(Y_i : i \in \mathcal{T})$, where ρ is an arbitrary permutation (re-ordering) of the treated individuals.

(ii). *If*, moreover, I were to take the aspirins, then the above exchangeability would extend to the set $\mathcal{T}^+ := \mathcal{T} \cup \{0\}$, in which I too am included.

Parallel exchangeability assumptions would be made for the control group $\mathcal{C}$, from whom the aspirin is withheld: in (i) and (ii) we just replace "treatment" by "control", $\mathcal{T}$ by $\mathcal{C}$ (and $\mathcal{T}^+$ by $\mathcal{C}^+$), and "were to take" by "were not to take". We shall denote these variant versions by (i)$'$ and (ii)$'$.

Since the above exchangeability assumptions relate to the responses of individuals after they have (actually or hypothetically) received treatment, we refer to them as *post-treatment exchangeabiity*.

Applying de Finetti's representation theorem (de Finetti 1937) to (i), I can regard the responses $(Y_i : i \in \mathcal{T})$ in the treatment group as independently and identically distributed, from some unknown distribution.[6] This distribution can then be consistently estimated from the response data in the treatment group. On account of (ii), this same distribution would govern my own response, Y_0, were I to take the aspirins. It can thus be identified with my own hypothetical distribution P_1. Taken together, (i) and (ii) thus justify my estimating P_1 from the treatment group data, and using this to populate the treatment branch of my decision tree.[7] Similarly, using (i)$'$ and (ii)$'$, I can use the data from the control group to populate my own control branch. My decision problem can now be solved.[8]

---

[6]Strictly, this result requires that I could, at least in principle, extend the size of the treatment group indefinitely, while retaining exchangeability.

[7]More correctly, I should take account of all the data, in both groups. I regard the associated ordered outcomes as *partially exchangeable* (de Finetti 1980), with a joint distribution unchanged under arbitrary permutations of individuals within each group. Such a joint distribution can be regarded as generated by independent sampling, from a distribution P_1 for an individual in the treatment group, or P_0 for an individual in the treatment group, where I have a joint distribution for the pair (P_0, P_1). There could be dependence between P_0 and P_1 in this joint distribution (for example, they might contain common parameters)—in which case data on responses in the control group could also carry information about the treatment response distribution P_1. Nevertheless, if the treatment data are sufficiently extensive I can still estimate P_1 consistently by ignoring the control data, and so use just the treatment data to populate the treatment arm of my decision problem.

[8] The above argument glosses over a small philosophical problem: Can I justify equating the *hypothetical* uncertainty about the response Y, *were an individual to take* the aspirins, with the *realised* uncertainty about (still unobserved) Y, once *that individual is known to have taken* the aspirins? (and, importantly, nothing else new is known). The former is what is relevant to my decision problem, but the data on the treated individuals are informative about the latter. We have implicitly assumed that these uncertainties are the same, and so governed by the same

Some comments

(1). Whether or not the exchangeability assumption (i) can be regarded as reasonable will be highly dependent on the background information informing my personal probability assessments. For example, I might know, or suspect, that evening headaches tend to be more long-lasting than morning headaches. If I were also to know which of the headaches in $\mathcal{T}$ were evening, and which morning, headaches, then I would not wish to impose exchangeability. I might know that individual 1 had a morning headache, and individual 2 an evening headache. Then it would not be reasonable for me to give the re-ordered pair (Y_2, Y_1) the same joint distribution as (Y_1, Y_2)—in particular, my marginal distribution for Y_2 would likely not be the same as that for Y_1. However, in the absence of specific knowledge about who had what type of headache— "equality of ignorance"—the exchangeability condition (i) could still be reasonable.

(2). There may be more than one way of embedding my own response, Y_0, into a set of exchangeable variables. For example, instead of considering other individuals, I could consider all my own previous headache episodes. (In the language of experimental design, the experimental unit—the headache episode—is nested within the individual). Then I might use the estimated distribution of my response, among those past headache episodes of my own that I had treated with aspirin, to populate the treatment branch of my current decision problem. This might well yield a different (and arguably more relevant) distribution from that based on observing headaches in other treated individuals. In this sense there is no "objective" distribution P_1 waiting to be uncovered: P_1 is itself an artifact of the overall structure in which I have embedded my problem, and the data that I have observed.

(3). Exchangeability must also be considered in relation to my own current circumstances. The exchangeability judgment (i) may not be extendible as required by (ii) if, for example, my current headache is particularly severe. To reinstate exchangeability I might then need to restrict attention to those headache episodes (in other individuals, or in my own past) that had a similar level of severity to mine. Alternatively I might build a more complex statistical model, allowing for different degrees of severity, and use this to extrapolate from the observed data to my own case.

(4). We do not in principle exclude complicated scenarios such as "herd immunity" in vaccination programmes, where an individual's response might be affected in part by the treatments that are assigned to other individuals. Assuming appropriate symmetry in (my knowledge of) the interactions between individuals, this need not negate the appropriateness of the exchangeability assumptions, and hence the validity of the above analysis—though in this case it would be difficult to give the underlying distributions P_0 and P_1, conjured into existence by de Finetti's theorem, a clear frequentist interpretation. However, in such a problem it would usually be more appropriate to enter into a more detailed modelling of the situation.

Exchangeability, while an enormously simplifying assumption, is in any case inessential for the more general analysis of § 5.2: at that level of generality, I have to assess my conditional distribution

---

distribution. We may term this property *temporal coherence*. At a fully general level, any conditional probability $P(A \mid B)$ has two different interpretations: the (hypothetical) probability it would be appropriate to assign to A, were B (and only B) to become known, and the (realised) probability it is appropriate to assign to A, after B (but nothing else new) has become known. Although it seems innocuous to equate these two, a full philosophical justification is not entirely trivial (see for example Skyrms (1987)). Nevertheless there is no serious dissent from this position, and we shall adopt it without further ado.

for my own response Y_0 (in the hypothetical situation that I decide to take the aspirins), given whatever data D I have available. But modelling and implementing an unstructured prediction problem can be extremely challenging, as well as hard to justify as genuinely empirically based, unless we can make good arguments. When appropriate, judgments of exchangeability constitute an excellent basis for such arguments.

6.2 Pre-treatment exchangeability

The post-treatment exchangeability conditions (i) and (ii), and (i)′ and (ii)′, are what is needed to let me populate my decision tree with the requisite hypothetical distributions and so solve my decision problem.

Here we consider another interpretation of the expression "a group of individuals whom I can regard, in an intuitive sense, as similar to myself, with headaches similar to my own". This description has been supposed equally applicable to the treatment group $\mathcal{T}$ and the control group $\mathcal{C}$. But this being the case, then—applying Euclid's first axiom, "Things which are equal to the same thing are also equal to one another"—the two groups, $\mathcal{T}$ and $\mathcal{C}$ (and their headaches), both being similar to me, must be regarded (again in an intuitive sense) as similar to each other—I must be "comparing like with like". But how are we to formalise this intuitive property of the two groups being similar to each other? We cannot simply impose full exchangeability of all the responses $(Y_i : i \in \mathcal{D})$, since I typically would not expect the responses of the treated individuals to be exchangeable with those of the untreated individuals.

One way of formalising this intuition is to consider all the individuals in the treatment and control groups *before* they were given their treatments. Just as I myself can hypothesise taking either one of the treatments, and in either case consider my hypothetical distribution for my ensuing response Y_0, so can I hypothesise various ways in which treatments might be applied to all the individuals in $\mathcal{I}$.

Let the binary decision variable $\check{T}_i$ indicate which treatment is hypothesised to be applied to individual i.

We first introduce the following *Stable Unit-Treatment Distribution Assumption*:

Condition 1 (SUTDA) *For any $A \subseteq \mathcal{I}$, the joint distribution of $Y_A := (Y_i : Y \in A)$, given hypothesised treatment applications $(\check{T}_i = t_i : i \in \mathcal{I})$, depends only on $(t_i : i \in A)$. In particular, for any individual i, the distribution of the associated response Y_i depends only on the treatment t_i applied to that individual.*

As discussed further in § 11.1 below, SUTDA bears a close resemblance to the Stable Unit-Treatment *Value* Assumption, SUTVA, typically made in the Rubin potential outcome framework; but—as reflected in it name—differs in the important respect of referring to distributions, rather than values, of variables. It is a weaker requirement than SUTVA, but is as powerful as required for applications.

Note that SUDTA is a genuinely restrictive hypothesis, now excluding cases such as the vaccine example (4) of § 6. However, we will henceforth assume it holds.

In more complex problems there will be other generic variables of interest besides Y—we term these (including the response variable Y) *domain variables*. Then we extend SUTDA to apply to all domain variables, considered jointly. An important special case is that of a domain variable X such that the joint distribution of $(X_i : i \in \mathcal{I})$, given $\check{T}_i = t_i$ $(i \in \mathcal{I})$, does not depend in any way on the applied treatments (t_i). Such a variable, unaffected by the treatment, is a *concomitant*. It will typically be reasonable to treat as a concomitant any variable whose value is fully determined before

the treatment decision has to be made: such a variable is termed a *covariate*. Other concomitants might include, for example, the weather after the treatment decision is made.

Let V be a (possibly multivariate) generic variable. I now hypothesise giving *all* individuals in $\mathcal{I}$ (including myself) the aspirins, and consider my corresponding hypothetical joint distribution for the individual instances $(V_i : i \in \mathcal{I})$. It would often be reasonable to impose full exchangeability on this joint distribution, since all members of $\mathcal{I}$ would have been treated the same. A similar assumption can be made for the case that the aspirins are, hypothetically, withheld from all individuals. We term the conjunction of these two hypothetical exchangeability properties *pre-treatment exchangeability* (of V, over $\mathcal{I}$).

When I can assume this, then under uniform application of aspirin, by de Finetti's theorem I can regard all the (V_i) as independent and identically distributed from some distribution Q_1 (initially unknown, but estimable from data on uniformly treated individuals). Similarly, under hypothetical uniform withholding of aspirin, there will be an associated distribution Q_0. When moreover SUTDA applies, we can conclude that, under any hypothesised application of treatments, $\check{T}_i = t_i$ ($i \in \mathcal{I}$), we can regard the V_i as independent, with $V_i \sim Q_{t_i}$. We can thus confine attention to the generic variable V, with distribution Q_1 [resp., Q_0] under applied treatment $\check{T} = 1$ [resp., $\check{T} = 0$].

Pre-treatment exchangeability appears, superficially, to be a stronger requirement than post-treatment exchangeability: one could argue that (taken together with SUDTA) pre-treatment exchangeability implies the post-treatment exchangeability properties (i), (ii), (i)$'$ and (ii)$'$, which would permit me to populate both the treatment and the control branches of my decision tree, and so solve my decision problem. This would indeed be so if the individuals forming the treatment and control groups were identified in advance, and then subjected to their appointed interventions. However, it need not be so in the more general case that we do not have direct control over who gets which treatment. Much of the rest of this paper is concerned with addressing such cases, considering further conditions—in particular, *ignorability* of the treatment assignment process, as described on § 7.1 below—that allow us to bridge the gap between pre- and post-treatment exchangeability.

6.3   Internal and external validity

We might be willing to accept pre-treatment exchangeability, but only over the restricted set $\mathcal{D}$ of data individuals, excluding myself—a property we term *internal exchangeability*. When I can extend this to pre-treatment exchangeability over the set $\mathcal{D}^+ := \mathcal{D} \cup \{0\}$, including myself, we have *external exchangeability*. In the latter case there is at least a chance that the data $\mathcal{D}$ could help me solve my decision problem—the case of *external validity* of the data.[9] However, when we have internal but not external exchangeability, this conclusion could, at best, be regarded as holding for a new, possibly fictitious, individual who could be regarded as exchangeable with those in the data—this is the case of *internal validity*. In practice that can be problematic. For example, a clinical trial might have tightly restricted enrolment criteria, perhaps restricting entry to, say, men aged between 25 and 49 with mild headache. Even if the study has good internal validity, and shows a clear advantage to aspirin for curing the headache, it is not clear that this message would be relevant to a 20-year old female with a severe headache. And indeed, it may not be. Arguments for external validity will generally be somewhat speculative, and not easy to support with empirical evidence.

---

[9]This is admittedly a very strict interpretation of "external validity". More generally, it might be considered enough to be able to transfer information about, say, ACE, from the data to me. This would typically require further modelling assumptions, such as described in §8.1 of Dawid (2000).

7 Treatment assignment and application

In §5.2 we talked in terms of identifying, quite separately, two groups of individuals, in each case supposed suitably exchangeable (both internally, and with me), where one of the groups is made to take, and the other made not to take, the aspirins. But typically the process is reversed: a single group of individuals, $\mathcal{D}$ say, is gathered, some of whom are then chosen to receive active treatment—thus forming the treatment group $\mathcal{T}$—with the remainder forming the control group $\mathcal{C}$.

In this case the treatment process has three stages:

(1). First, the data subjects $\mathcal{D}$ are identified by some process.

(2). Secondly, certain individuals in $\mathcal{D}$ are somehow selected to receive active treatment, the others receiving control.[10]

(3). Finally, the assigned treatments are actually administered.

The operation of stage (1) will be crucial for issues of external validity—if the data are to be at all relevant for me, I would want the data subjects to be somehow like me. However from this point on we shall naïvely assume this has been done satisfactorily—alternatively, we consider "me" to be a possibly fictitious individual who can be regarded as similar to those in the data. We shall thus consider all data subjects, together with myself, as pre-treatment exchangeable. I can then confine attention to the joint distributions P_1 and P_0 over generic variables, under hypothesised application of treatment 1 or 0, respectively.

For further analysis it will prove important to keep stages (2) and (3) clearly distinct in the notation and the analysis.

We denote by T^* the generic *intention to treat* (ITT) variable, generated at stage (2), where $T_i^* = 1$ if individual $i \in \mathcal{D}$ is selected to receive active treatment, and $T_i^* = 0$ if not (this is relevant only for the external data $\mathcal{D}$: my own value T_0^* need not be defined). Note that T^* is a stochastic variable. In contrast, we also consider (at stage (3)) the binary non-stochastic generic decision/regime variable $\check{T}$: $\check{T}_i = 1$ [resp., $\check{T}_i = 0$] denotes the (typically hypothetical) situation in which individual i is made to take [resp., prevented from taking] the aspirins. My own decision variable $\check{T}_0$ (though not yet its value) is well-defined—indeed, is the very focus of my decision problem.

Note that when below we talk of "domain variables" we will exclude T^* and $\check{T}$ from this description.

If all goes to plan, for $i \in \mathcal{D}$ we shall have $\check{T}_i = T_i^*$. However, there is no bar to considering, between stages (2) and (3), what might happen to an individual, fingered to receive the treatment (so having $T_i^* = 1$), who, contrary to plan, is prevented from taking it (so that $\check{T}_i = 0$)[11]—indeed,

---

[10]In reality stages (1) and (2) may be combined, as in sequential accrual and randomisation in a clinical trial.

[11]This apparently oxymoronic combination has some superficial resemblance to counterfactual reasoning (see *e.g.* Morgan and Winship (2014)), which has often been considered—quite wrongly in my view (Dawid 2000)—as essential for modelling and manipulating causal relations. Counterfactual analysis considers the individual after he has been treated (so with known $\check{T}_i = 1$, and possibly known response Y_i), and asks what might have happened if—in a fictional scenario *counter to known facts*—he had not been treated (*i.e.*, under the counterfactual application $\check{T}_i = 0$). In spite of some parallels, there are important differences between our *hypothetical* approach and this *counterfactual* approach. By considering a time before any treatment has yet been applied, and making the distinction between intention to treat, T_i^*, and a hypothesised treatment application, $\check{T}_i$, we sidestep many of the philosophical and methodological difficulties associated with counterfactual reasoning. In particular, in our formulation we avoid counterfactual theory's problematic and entirely unnecessary conversion of the single response variable Y into two separate but co-existing "potential responses", $Y(0)$ and $Y(1)$.

we have already made use of such considerations when introducing pre-treatment exchangeability. So we can meaningfully consider a quantity such as $\mathrm{E}(Y \mid T^* = 1, \check{T} = 0)$. And indeed it will prove useful to divorce treatment *selection* (intention to treat), T^*, from (actual or hypothetical) treatment *application*, $\check{T}$, in this way. For example, what is usually termed *the effect of treatment on the treated* (Heckman 1992) is more properly expressed as *the effect of treatment on those selected for treatment*, which can be represented formally as $\mathrm{E}(Y \mid T^* = 1, \check{T} = 1) - \mathrm{E}(Y \mid T^* = 1, \check{T} = 0)$ (Geneletti and Dawid 2011).

Since the selection process is made before any application of treatment, it is appropriate to treat T^* as a covariate, with the same distribution in both regimes.

We suppose internal exchangeability, in the sense of §6.2 above, for the pair of generic variables (T^*, Y). In particular we shall have internal exchangeability, marginally, for the response variable Y—and, to make a link to my own decision problem, we assume this extends to external exchangeability for Y (we here omit T^*, since that might not even be meaningfully defined for me). However, even internal exchangeability for Y need no longer hold after we condition on the selection variable T^*—this is the problem of *confounding*. For example, suppose that, although I myself don't know which of the headaches in $\mathcal{D}$ are the (generally milder) morning and which the (generally more long-lasting) evening headaches, I know or suspect that the aspirins have been assigned preferentially to the evening headaches. Then simply knowing that an individual was selected (perhaps self-selected) to take the aspirins ($T^* = 1$) will suggest that his headache is more likely to be an evening headache, and so change my uncertainty about his response Y (whichever treatment were to be taken). I might thus expect, *e.g.*, $\mathrm{E}(Y \mid T^* = 1, \check{T} = t) > \mathrm{E}(Y \mid T^* = 0, \check{T} = t)$, both for $t = 0$ and for $t = 1$. In such a case, even under a hypothetical uniform *application* of treatment, I could not reasonably assume exchangeability between the group *selected* to receive active treatment (and thus more likely to have long-lasting evening headaches) and the group selected for control (who are more likely to have short-lived morning headaches). Post-treatment exchangeability is absent, since I would no longer be comparing like with like. This in turn renders external validity impossible, since (even under uniform treatment) I could not now be exchangeable, simultaneously, both with those selected for treatment and with the those selected for control, since these are not even exchangeable with each other. This means I can no longer use the data (at any rate, not in the simple way considered thus far) to fully populate, and thus solve, my decision problem.

As explained in §6.2, assuming internal exchangeability and SUTDA, I can just consider the joint distribution, Q_t, for the bivariate generic variable (T^*, Y), given $\check{T} = t$. Since we are treating the selection indicator T^* as a covariate, its marginal distribution will not depend on which hypothetical treatment application is under consideration, and so will be the same under both Q_1 and Q_0. We can express this as the extended independence property

$$T^* \perp\!\!\!\perp \check{T}, \tag{8}$$

which says that the (stochastic) selection variable T^* is independent of the (non-stochastic) decision variable $\check{T}$. We denote this common distribution of T^* in both regimes by P^*.

By the assumed external exchangeability of Y, the marginal distribution of Y under Q_t is my desired hypothetical response distribution, P_t. However, in the absence of actual uniform application of treatment t to the data subjects (which in any case is not simultaneously possible for both values of t), I may not be able to estimate this marginal distribution. In the data, the treatment will have been applied in accordance with the selection process, so that $\check{T} = T^*$, and the only observations I will have under regime $\check{T} = 1$ (say) are those for which $T^* = 1$. From these I can estimate the *conditional* distribution of Y, given $T^* = 1$. under Q_1—but this need not agree

with the desired *marginal* distribution P_1 of Y under Q_1.[12]

7.1 Ignorability

The above complication will be avoided when I judge that, both for $t = 1$ and for $t = 0$, if I intervene to *apply* treatment $\check{T} = t$ on an individual, the ensuing response Y will not depend on the *intended* treatment T^* for that individual; *i.e.*, we have independence of Y and T^* under each Q_t. This can be expressed as the extended conditional independence property

$$Y \perp\!\!\!\perp T^* \mid \check{T}. \tag{9}$$

When (9) can be assumed to hold, we term the assignment process *ignorable*. In that case, my desired distribution for Y, under hypothesised active treatment assigment $\check{T} = 1$, is the same as the conditional distribution of Y given $T^* = 1$ under $\check{T} = 1$—which is estimable as the distribution of Y in the treatment group data. Likewise, my distribution for Y under hypothesised control treatment is estimable from the data in the control group.

The ignorability condition (9) requires that the distribution of an individual's response Y, under either applied treatment, will not be affected by knowledge of which treatment the individual had been fingered to receive—a property that would likely fail if, for example, treatment selection T^* was related to the overall health of the patient. Note that ignorability is not testable from the available data, in which $\check{T} = T^*$. For we would need to test, in particular, that, for an individual taking actual treatment $\check{T} = 1$, the distribution of Y given $T^* = 1$ is the same as that given $T^* = 0$. But for all such individuals in the data we never have $T^* = 0$, so can not make the comparison. Hence any assumption of ignorability can only be justified on the basis of non-empirical considerations. The most common, and most convincing, basis for such a justification is when I know that the treatment assignment process has been carried out by a randomising device, which can be assumed to be entirely unrelated to anything that could affect the responses; but I might be able to make a non-empirical arguments for ignorability in some other contexts also. Indeed, it would be rash simply to assume ignorability without having a good argument to back it up.

8 The idle regime

As a useful extension of the above analysis, we expand the range of the regime indicator $\check{T}$ to encompass a further value, which we term "idle", and denote by $\emptyset$—this indicates the observational regime, where treatments are applied according to plan. (This is relevant only for the data individuals, in $\mathcal{D}$: I myself care only about the two interventions I am considering). We denote this 3-valued regime indicator by F_T.

Now T^* is determined prior to any (actual or hypothetical) treatment application, and behaves as a covariate. It is thus reasonable to assume that, under the observational regime $F_T = \emptyset$, T^* retains its fixed covariate distribution P^*. And since this distribution is then the same in all three

---

[12]Note that I can not make use of the *conditional* distributions of Y given T^*. Typically I myself do not even have, let alone observe, a value for T_0^*. And even in the special case that my value T_0^* is well-defined, and I can assume external validity for the pair (T^*, Y), I can at best estimate one of the two required conditional distributions. Thus if I have been fingered for treatment, $T_0^* = 1$, I would need the conditional distribution of Y given $T^* = 1$ under Q_t, for both $t = 0$ and $t = 1$. But for $t = 0$ this will not be estimable from the data, since there were no data subjects who were fingered for treatment but did not receive it.

regimes, we thus have

$$T^* \perp\!\!\!\perp F_T. \tag{10}$$

This extends (8) to include also the idle regime.

We now introduce a new stochastic domain variable T, representing the treatment actually applied when following the relevant regime. This is fully determined by the pair (F_T, T^*), as follows:

Definition 1 (Applied Treatment, T)

  (i). If $F_T = 0$ or 1, then $T = F_T$

 (ii). If $F_T = \emptyset$, then $T = T^*$.

In particular, $T \sim P^*$ under $F_T = \emptyset$, while T has a degenerate distribution at t under $F_T = t$ ($t = 0$ or 1). □

In each of the three regimes we can observe both T and Y. In the observational regime ($F_T = \emptyset$) we can also recover T^*, since $T^* = T$. However, T^* is typically unobservable in the interventional regimes, and may not even be defined for myself, the case of interest.

To complete the distributional specification of the idle regime we argue as follows. Under $F_T = \emptyset$, the information conveyed by learning $T = t$ is twofold, conveying both that the individual was initially fingered to receive treatment t, *i.e.* $T^* = t$, and that treatment t was indeed applied. Hence for any domain variable V, the conditional distribution of V given $T = t$ (equivalently, given $T^* = t$), under $F_T = \emptyset$, should be the same as that of V given $T^* = t$, under the (real or hypothetical) applied treatment $F_T = t$. We express this property formally as:

Definition 2 (Distributional Consistency) For any domain variable, or set of domain variables, V,[13]

$$V \mid (T = t, F_T = \emptyset) \; [= V \mid (T^* = t, F_T = \emptyset)] \; \approx \; V \mid (T^* = t, F_T = t) \qquad (t = 0, 1), \tag{11}$$

where $\approx$ denotes "has the same distribution as". □

Distributional consistency is the fundamental property linking the observational and interventional regimes. It is our, weaker, version of the (functional) consistency property usually invoked in the potential outcome approach to causality—see §11.1 below. In the sequel we shall take (11) for granted.

Lemma 1 *For any domain variable V,*

$$V \perp\!\!\!\perp F_T \mid (T, T^*). \tag{12}$$

Proof.   We have to show that, for $t, t^* \in \{0, 1\}$, it is possible to define a conditional distribution for V, given $T = t, T^* = t^*$, that applies in all three regimes.

Let Π_{t,t^*} denote the distribution of V given $T^* = t^*$ in the interventional regime $F_T = t$. This is well-defined in the usual case that the event $T^* = t^*$ has positive probability (this probability being the same in all regimes)—if not, we make an arbitrary choice for this distribution.

Consider first the case $t = 1$.

---

[13]Note that (11) holds, automatically, on taking for V the constructed variable T, since on each side the conditional distribution for $V \equiv T$ is the one-point distribution on the value t.

18

(1). Since T is non-random with value 1 in regime $F_t = 1$, Π_{1,t^*} is also, trivially, the distribution of V given $T = 1, T^* = t^*$ in regime $F_T = 1$.

(2). Under regime $F_T = 0$, the event $T = 1, T^* = t^*$ has probability 0, so we are free to define the distribution of V conditional on this event arbitrarily; in particular we can take it to be Π_{1,t^*}.

(3). Under regime $F_T = \emptyset$, the event $T = 1, T^* = 0$ has probability 0, so we are free to define the distribution of V conditional on this event as $\Pi_{1,0}$.

(4). It remains to show that the distribution of V given $T = T^* = 1$ in regime $F_T = \emptyset$ is $\Pi_{1,1}$. Since, under $F_T = \emptyset$, $T \equiv T^*$, we need only condition on $T = 1$. The result now follows from distributional consistency (11).

Since a parallel argument holds for the case $t = 0$, we have shown that Π_{t,t^*} serves as the conditional distribution for V given $(T = t, T^* = t^*)$ in all three regimes, and (12) is thus proved.
□

8.1 Graphical representation

The properties (10) and (12) are represented graphically (using d-separation) by the absence of arrows from F_T to T^* and to Y, respectively, in the ITT (intention to treat) DAG of Figure 5, where again, a round node represents a stochastic variable, and a square node a non-stochastic regime indicator. In addition, we have included further optional annotations:

- The outline of T^* is dotted to indicate that T^* is not directly observed

- The heavy outline of T indicates that the value of T is *functionally* determined by those of its parents F_T and T^*

- The dashed arrow from T^* to T indicates that this arrow can be removed (there is then no dependence of T on T^*) under either of the *interventional* settings $F_T = 0$ or 1.
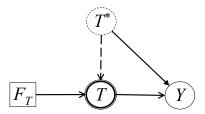

Figure 5: DAG representing $T^* \perp\!\!\!\perp F_T$ and $Y \perp\!\!\!\perp F_T \mid (T, T^*)$

Remark 1 Note that, on further taking into account the functional relationship of Definition 1, Figure 5 already incorporates the distributional consistency property of Definition 2, for $V \equiv Y$. For we have

$$
\begin{align}
Y \mid (T = t, F_T = \emptyset) &= Y \mid (T = t, T^* = t, F_T = \emptyset) \tag{13} \\
&\approx Y \mid (T = t, T^* = t, F_T = t) \tag{14} \\
&= Y \mid (T^* = t, F_T = t). \tag{15}
\end{align}
$$

Here (13) follows from (ii) of Definition 1; (14) from Lemma 1 with $V \equiv Y$, *i.e.* $Y \perp\!\!\!\perp F_T \mid (T, T^*)$, which is represented in Figure 5; and (13) from (i) of Definition 1. $\square$

Now the ITT variable T^*, while crucial to understanding the relationship between the different regimes, is not itself directly observable. If we confine attention to relationships between the domain variables, Figure 5 collapses into the essentially vacuous DAG of Figure 6, expressing no non-trivial conditional independence properties. So without further assumptions there is no useful structure
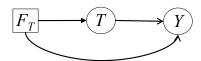

Figure 6: Vacuous DAG between F_T, T and Y, ignoring T^*

of which to avail ourselves.

8.2 Ignorability

Suppose now we impose the additional ignorability property (9). Noting that $\check{T} = t$ is identical with $F_T = t$, this is equivalent to

$$Y \perp\!\!\!\perp T^* \mid F_T = t \qquad (t = 0, 1). \tag{16}$$

Equivalently, since T is non-random in an interventional regime,

$$Y \perp\!\!\!\perp T^* \mid (T, F_T = t) \qquad (t = 0, 1).$$

Moreover, since in the idle regime, T^* is identical with T, so non-random when T is given, we trivially have

$$Y \perp\!\!\!\perp T^* \mid (T, F_T = \emptyset).$$

We thus see that ignorability can be expressed as:

$$Y \perp\!\!\!\perp T^* \mid (T, F_T). \tag{17}$$

Lemma 2 *If ignorability holds, then*

$$Y \perp\!\!\!\perp F_T \mid T. \tag{18}$$

Proof. We first dispose of the trivial case that T^* has a one-point distribution. In that case the conditioning on T^* in (12) is redundant and we immediately obtain (18).

Otherwise, $0 < \mathrm{pr}(T^* = 1) < 1$. We then have

$$
\begin{align}
Y \mid (T = 1, F_T = \emptyset) \quad &\approx \quad Y \mid (T^* = 1, F_T = 1) \tag{19} \\
&\approx \quad Y \mid F_T = 1 \tag{20} \\
&\approx \quad Y \mid (T = 1, F_T = 1). \tag{21}
\end{align}
$$

Note that all conditioning events have positive probability in their respective regimes. Here (19) holds by distributional consistency (11), (20) by ignorability (16), and (21) because, under $F_T = 1$, $T = 1$ with probability 1. So we have a common well-defined distribution, Δ_1 say, for Y given $T = 1$ in both regimes $F_T = \emptyset$ and $F_T = 1$. Further, since under $F_T = 0$ the event $T = 1$ has probability 0, we are free to define the conditional distribution of Y given $T = 1$ in regime $F_T = 0$ as Δ_1 also, so making Δ_1 the common distribution of Y given $T = 1$ in all 3 regimes, showing that $Y \perp\!\!\!\perp F_T \mid T = 1$. Since a similar argument holds for conditioning on $T = 0$ the result follows. $\quad\square$

Remark 2 An apparently simpler alternative proof of Lemma 2 is as follows. By Lemma 1, the conditional distribution of Y, given (F_T, T, T^*), does not depend on F_T, while by (17) this conditional distribution does not depend on T^*. So (it appears), it must follow that it depends only on T, whence $Y \perp\!\!\!\perp (F_T, T^*) \mid T$, implying the desired result. This is a special case of a more general argument: that $X \perp\!\!\!\perp Y \mid (Z, W)$ and $X \perp\!\!\!\perp Z \mid (Y, W)$ together imply $X \perp\!\!\!\perp (Y, Z) \mid W$. However this argument is invalid in general (Dawid 1979b). To justify it in this case we have needed, in our proof Lemma 2, to call on structural properties (in particular, distributional consistency, and the way in which T is determined by F_T and T^*) in addition to conditional independence properties. $\quad\square$

Corollary 1 *Ignorability holds if and only if*

$$
Y \perp\!\!\!\perp (T^*, F_T) \mid T. \tag{22}
$$

Proof.

If: Further conditioning (22) on F_T yields (17).

Only if: Property (22) is equivalent to the conjunction of (17) and (18).

$\quad\square$

8.2.1 Graphical representation

The DAG representing (10) and (22) is shown in Figure 7. Compared with Figure 5, we see that the arrow from T^* to Y has been removed.

Remark 3 We might try and make the deletion of the arrow from T^* to Y in Figure 5 into a graphically based argument for Lemma 2, for it appears to impose just the additional conditional independence property (17) representing ignorability, and to imply the desired result (18). However this is again a misleading argument: inference from such surgery on a DAG can only be justified when it has a basis in the algebraic theory of conditional independence (Dawid 1979a; Constantinou and Dawid 2017), which here it does not, on account of the fallacious argument identified in Remark 2. $\quad\square$
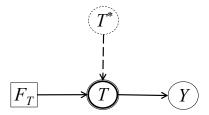
Figure 7: Modification of Figure 5 representing ignorability

Figure 8 results on "eliminating T^*" from Figure 7: that is to say, the conditional independencies represented in Figure 8 are exactly those of Figure 7 that do not involve T^*. In this case, the only such property is (18).

Figure 8: Collapsed DAG under ignorability, representing $Y \perp\!\!\!\perp F_T \mid T$

The ECI property (18), and the DAG of Figure 8, are the basic (respectively algebraic and graphical) representations of "no confounding" in the DT approach, which has been treated as a primitive in earlier work. The above analysis supplies deeper understanding of these representations. Although on getting to this point we have been able to eliminate explicit consideration of the treatment selection variable T^*, our more detailed analysis, which takes it into account, makes clear just what needs to be argued in order to justify (18): namely, the property of ignorability expressed algebraically by (16) or (17) and graphically by Figure 7, and further described in § 7.1.

9 Covariates

The ignorability assumption (9) will often be untenable. If, for example, those fingered for treatment (so with $T^* = 1$) are sicker than those fingered for control ($T^* = 0$)—as might well be the case in a non-randomised study—then (under either treatment application $\tilde{T} = t$, $t = 0, 1$) we would expect a worse outcome Y when knowing $T^* = 1$ than when knowing $T^* = 0$. However, we might be able to reinstate (9) after further conditioning on a suitable variable X measuring how sick an individual is. That is, we might be able to make a case that, *after restricting attention to those individuals having a specified degree $X = x$ of sickness*, the further information that an individual had been fingered for treatment would make no difference to the assessment of the individual's response (under either treatment application). This would of course require that, after taking sickness into account, the treatment assignment process was not further related to other possible indicators of outcome (*e.g.*, sex, age,...). If it is, these would need to be included as components of the (typically multivariate) variable X. We assume that the appropriate variable X is (in principle at least) fully measurable, both for the individuals in the study and (unlike T^*) for myself. We assume internal exchangeability

of (X, T^*, Y), extending this to external exchangeability for (X, Y).[14]

If and when such a variable X can be identified, we will be able to justify an assumption of *conditional ignorability*:

$$Y \perp\!\!\!\perp T^* \mid (X, \check{T}). \tag{23}$$

Furthermore, to be of any use in addressing my own decision problem, such a variable must be a covariate, available prior to treatment application, and so, in particular must (jointly with T^*, at least for the study individuals, for whom T^* is defined) have the same distribution under either hypothetical treatment application. This is expressed as

$$(X, T^*) \perp\!\!\!\perp \check{T}. \tag{24}$$

In particular, there will be a common marginal distribution, P_X say, for X, in both interventional regimes.

When both (23) and (24) are satisfied, we call X a *sufficient covariate*. These properties are represented by the DAG of Figure 9.
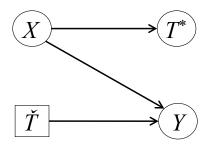

Figure 9: DAG representing sufficient covariate X: $(X, T^*) \perp\!\!\!\perp \check{T}$ and $Y \perp\!\!\!\perp T^* \mid (X, \check{T})$.

9.1   Idle regime

As in §8, we introduce the regime indicator F_T, allowing for consideration of the "idle" observational regime $F_T = \emptyset$, in addition to the interventional regimes $F_T = t$ ($t = 0, 1$); and the constructed "applied treatment" variable T of Definition 1. Arguing as for (17), (23) implies

$$Y \perp\!\!\!\perp T^* \mid (X, T, F_T). \tag{25}$$

Lemma 3 *Let X be a sufficient covariate. Then*

$$
\begin{aligned}
(X, T^*) &\quad \perp\!\!\!\perp \quad F_T & (26) \\
Y &\quad \perp\!\!\!\perp \quad (T^*, F_T) \mid (X, T). & (27)
\end{aligned}
$$

Proof.   By distributional consistency (11),

$$
\begin{aligned}
X \mid T^* = 1, F_T = \emptyset \quad &\approx \quad X \mid T^* = 1, F_T = 1 \\
&\approx \quad X \mid T^* = 1, F_T = 0
\end{aligned}
$$

---

[14]This last condition could be relaxed, allowing my own distribution for X to differ from that in the data, while retaining conditional exchangeability for Y, given X. For simplicity we do not consider this further here.

by (24). Hence $X \perp\!\!\!\perp F_T \mid T^* = 1$. A parallel argument shows $X \perp\!\!\!\perp F_T \mid T^* = 0$, so that $X \perp\!\!\!\perp F_T \mid T^*$. On combining this with (10) we obtain (26).

As for (27), this is equivalent to the conjunction of (25) and $Y \perp\!\!\!\perp F_T \mid (T, X)$. The argument for the latter (again, requiring distributional consistency) parallels that for (18), after further conditioning on X throughout. $\qquad\square$

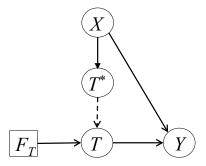The properties (26) and (27) are embodied in the DAG of Figure 10. This implies, on eliminating

Figure 10: Full DAG with sufficient covariate X and regime indicator

the unobserved variable T^*:

$$X \quad \perp\!\!\!\perp \quad F_T \tag{28}$$
$$Y \quad \perp\!\!\!\perp \quad F_T \mid (X, T), \tag{29}$$

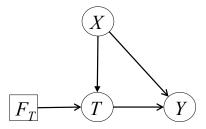as represented by Figure 11. Properties (28) and (29), as embodied in Figure 11, are the basic DT

Figure 11: Reduced DAG with sufficient covariate X and regime indicator

representations of a sufficient covariate. Assuming X, T and Y are all observed, this is what is commonly referred to as "no unmeasured confounding".

10 More complex DAG models

10.1 An example

Consider the following story. In an observational setting, variable X_0 represents the initial treatment received by a patient; this is supposed to be applied independently of an (unobserved) characteristic H of the patient. The variable Z is an observed response depending, probabilistically, on both the applied treatment X_0 and the patient characteristic H. A subsequent treatment, X_1, can depend probabilistically on both Z and H, but not further on X_0. Finally the distribution of the response Y, given all other variables, depends only on X_1 and Z. Figure 12 is a DAG representing this story by means of d-separation.
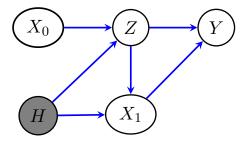

Figure 12: Observational DAG

In addition to the observational regime, we want to consider possible interventions to set values for X_0 and X_1. We thus have two non-stochastic regime indicators, F_0 and F_1: $F_i = x_i$ indicates that X_i is externally set to x_i, while $F_i = \emptyset$ allows X_i to develop "naturally". The overall regime is thus determined by the pair (F_0, F_1).

Figure 13 augments Figure 12, in a seemingly natural way, to include these regime indicators. It represents, by d-separation, ways in which the domain variables are supposed to respond to interventions. For example, it implies $Y \perp\!\!\!\perp (X_0, H, F_0, F_1) \mid (Z.X_1)$: once we know Z and X_1, not only are X_0 and H irrelevant for probabilistic prediction of Y, but so too is the information as to whether either or both of X_0, X_1 arose naturally, or were set by intervention. In particular, the conditional distribution of Y given (Z, X_1), under intervention at X_1, is supposed the same as in the observational regime modelled by Figure 12.

10.1.1 From observational to augmented DAG

It does not follow, merely from the fact that we can model the observational conditional independencies between the domain variables by Figure 12, that their behaviour under the entirely different circumstance of intervention must be as modelled by Figure 13. Strong additional assumptions are required to bridge this logical gap. These we now elaborate.

We again introduce "intention to treat" variables, X_0^* and X_1^*, the realised X_0 and X_1, in any regime, being given by

$$X_i = \begin{cases} X_i^* & \text{if } F_i = \emptyset \\ F_i & \text{if } F_i \neq \emptyset. \end{cases} \tag{30}$$
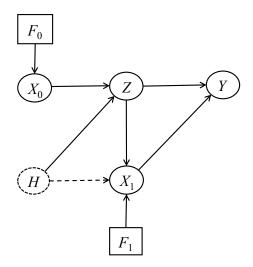
Figure 13: Augmented DAG

Since, in the observational regime, $X_i = X_i^*$, Figure 12 would still be observationally valid on replacing each X_i by X_i^*.

The different regimes are supposed linked together by the following assumptions, which we first present and then motivate:

$$X_0^* \quad \perp\!\!\!\perp \quad (F_0, F_1) \tag{31}$$

$$(H, Z, X_1^*, Y) \quad \perp\!\!\!\perp \quad (F_0, X_0^*) \mid (F_1, X_0) \tag{32}$$

$$(X_0^*, H, Z, X_1^*) \quad \perp\!\!\!\perp \quad F_1 \mid F_0 \tag{33}$$

$$Y \quad \perp\!\!\!\perp \quad (F_1, X_1^*) \mid (F_0, X_0, H, Z, X_1). \tag{34}$$

Note that, since X_i is determined by (F_i, X_i^*), (32) and (33) are equivalent to:

$$(H, Z, X_1^*, X_1, Y) \quad \perp\!\!\!\perp \quad (F_0, X_0^*) \mid (F_1, X_0) \tag{35}$$

$$(X_0^*, X_0, H, Z, X_1^*) \quad \perp\!\!\!\perp \quad F_1 \mid F_0. \tag{36}$$

Comments on the assumptions In order to understand the above assumptions, we should consider Figure 12 as describing, not only the conditional independencies between variables, but also a partial order in which the variables are generated: it is supposed that, in any regime, the value of a parent variable is determined before that of its child. In particular it is assumed that an intervention on a variable can not affect that variable's non-descendants—including their intention-to-treat variables and its own; but may affect its descendants—including their associated intention-to-treat variables.

(i). Similar to (10), (31) expresses the property that an intention-to-treat variable, here X_0^*, should behave as a covariate for X_0, and so be independent of which regime, here F_0, is operating on X_0. Moreover, X_0^* should not be affected by a subsequent intervention (or none), F_1, at X_1.

26

(ii). Assumption (32) is a version of the ignorability property (22). It says that an intervention on X_0 should be ignorable in its effect on all other variables. Moreover this should apply conditional on F_1, *i.e.* whether or not there is an intervention at X_1.

Remark 4 As previously discussed, ignorability is a strong assumption, requiring strong justification. Also note that, as shown by Corollary 1, (32) is implicitly assuming the distributional consistency property (Definition 2), in addition to ignorability. □

(iii). Assumption (33) expresses the requirement that (X_0^*, H, Z, X_1^*), being generated prior to X_1, should not be affected by intervention F_1 at X_1. (However, they might depend on which regime, F_0, operates on X_0.)

(iv). Similar to (ii), (34) says that, conditional on all the domain variables, (X_0, H, Z), generated prior to X_1, the effect of intervention F_1 at X_1 is ignorable for its effect on Y; moreover, this should hold whether or not there is intervention F_0 at X_0. Informally, taken together with (36), this requires that (X_0, H, Z) form a sufficient covariate for the effect of X_1 on Y.

In the following we make extensive (but largely implicit) use of the axiomatic properties of (extended) conditional independence (Dawid 1979a; Pearl 1988):

P1 (Symmetry): $X \perp\!\!\!\perp Y \mid Z \Rightarrow Y \perp\!\!\!\perp X \mid Z$.

P2: $X \perp\!\!\!\perp Y \mid Y$.

P3 (Decomposition): $X \perp\!\!\!\perp Y \mid Z$ and W a function of $Y \Rightarrow X \perp\!\!\!\perp W \mid Z$.

P4 (Weak Union): $X \perp\!\!\!\perp Y \mid Z$ and W a function of $Y \Rightarrow X \perp\!\!\!\perp Y \mid (W, Z)$.

P5 (Contraction): $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z) \Rightarrow X \perp\!\!\!\perp (Y, W) \mid Z$.

Lemma 4 *Suppose that the observational conditional independencies are represented by Figure 12, and that assumptions (31)–(34) apply. Then the extended conditional independencies between domain variables, intention-to-treat variables and regime indicators are represented by Figure 14.*
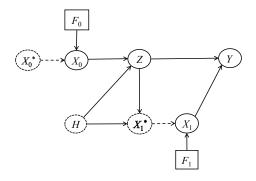

Figure 14: ITT DAG

Remark 5 A further property apparently represented in Figure 14 is the independence of F_0 and F_1:

$$F_0 \perp\!\!\!\perp F_1. \tag{37}$$

Now so far we have been able to meaningfully interpret an extended conditional independence assertion only when the left-hand term involves stochastic variables only—which seems to render (37) meaningless. Nevertheless, as a purely instrumental device, it is helpful to extend our understanding by considering the regime indicators as random variables also. So long as all our assumptions and conclusions do not involve regime indicators in their left-hand term, any proof that uses this extended understanding will remain valid for the actual case of non-stochastic regime variables, as may be seen by conditioning on these.[15]   □

In the light of Remark 5, we shall in the sequel treat F_0 and F_1 as stochastic variables, having the independence property (37).

Proof of Lemma 4.   Taking the variables in the order $F_0, F_1, X_0^*, X_0, H, Z, X_1^*, X_1, Y$, we need to show the following series of properties, where each asserts the independence of a variable from its predecessors, conditional on its parents in the graph.

$$
\begin{align}
F_1 \quad &\perp\!\!\!\perp \quad F_0 \tag{38}\\
X_0^* \quad &\perp\!\!\!\perp \quad (F_0, F_1) \tag{39}\\
X_0 \quad &\perp\!\!\!\perp \quad F_1 \mid (X_0^*, F_0) \tag{40}\\
H \quad &\perp\!\!\!\perp \quad (F_0, F_1, X_0^*, X_0) \tag{41}\\
Z \quad &\perp\!\!\!\perp \quad (F_0, F_1, X_0^*) \mid (X_0, H) \tag{42}\\
X_1^* \quad &\perp\!\!\!\perp \quad (F_0, F_1, X_0^*, X_0) \mid (H, Z) \tag{43}\\
X_1 \quad &\perp\!\!\!\perp \quad (F_0, X_0^*, X_0, H, Z) \mid (X_1^*, F_1) \tag{44}\\
Y \quad &\perp\!\!\!\perp \quad (F_0, F_1, X_0^*, X_0, H, X_1^*) \mid (Z, X_1). \tag{45}
\end{align}
$$

On excluding (38), these conclusions will comprise the desired result.

For (38): By assumption (37).

For (39): By (31).

For (40): Follows trivially since X_0, being functionally determined by (X_0^*, F_0), has a conditional one-point distribution, and so is independent of anything else.

For (41)–(43): From (35) we have

$$(H, Z, X_1^*) \perp\!\!\!\perp F_0 \mid (F_1, X_0) \tag{46}$$

while from (36) we have

$$(H, Z, X_1^*) \perp\!\!\!\perp F_1 \mid (F_0, X_0). \tag{47}$$

---

[15]It is not strictly necessary to regard the regime indicators as stochastic. Instead we can interpret (37) as expressing the non-stochastic property of *variation independence* (Dawid 2001), meaning that the range of possible values for each is unconstrained by the value taken by the other. Indeed, this is implicit in our interpretive comments (i)–(iv) on conditions (31)–(34). We can then combine these two distinct interpretations of independence within the same application, as we do here. For a rigorous analysis see Constantinou and Dawid (2017).

We now wish to show that (46) and (47) imply

$$(H, Z, X_1^*) \perp\!\!\!\perp (F_0, F_1) \mid X_0. \tag{48}$$

This requires some caution, on account of Remark 2. To proceed we use the fictitious independence property (37).

From (36) we have $X_0 \perp\!\!\!\perp F_1 \mid F_0$, which together with (37) yields $F_1 \perp\!\!\!\perp (F_0, X_0)$, so that

$$F_1 \perp\!\!\!\perp F_0 \mid X_0. \tag{49}$$

Combining (46) and (49) yields $(F_1, H, Z, X_1^*) \perp\!\!\!\perp F_0 \mid X_0$ whence

$$(H, Z, X_1^*) \perp\!\!\!\perp F_0 \mid X_0. \tag{50}$$

Finally, combining (50) and (47) yields (48).

Now (48) asserts that the conditional distribution of (H, Z, X_1^*) given X_0 is the same in all regimes. In particular (noting that $X_1^* = X_1$ in the observational regime), that conditional distribution inherits the independencies of Figure 12. Properties (41)–(43) follow (on noting that X_0, being a function of F_0 and X_0^*, is redundant in (41) and (43)).

For (44): Trivial since X_1 is functionally determined by (F_1, X_1^*).

For (45): From (35) we derive both

$$Y \quad \perp\!\!\!\perp \quad F_0 \mid (F_1, X_0, H, Z, X_1) \tag{51}$$
$$Y \quad \perp\!\!\!\perp \quad X_0^* \mid (F_0, F_1, X_0, H, Z, X_1^*, X_1). \tag{52}$$

while from (34) we have

$$Y \quad \perp\!\!\!\perp \quad F_1 \mid (F_0, X_0, H, Z, X_1) \tag{53}$$
$$Y \quad \perp\!\!\!\perp \quad X_1^* \mid (F_0, F_1, X_0, H, Z, X_1). \tag{54}$$

We first want to show that (51) and (53) are together equivalent to

$$Y \perp\!\!\!\perp (F_0, F_1) \mid (X_0, H, Z, X_1). \tag{55}$$

To work towards this, we note that, by (35), $(H, Z, X_1) \perp\!\!\!\perp F_0 \mid (F_1, X_0)$, which together with (49) gives $(F_1, H, Z, X_1) \perp\!\!\!\perp F_0 \mid X_0$, whence

$$F_0 \perp\!\!\!\perp F_1 \mid (X_0, H, Z, X_1). \tag{56}$$

Then (55) follows from (51), (53) and (56) in parallel to the argument above from (46), (47) and (49) to (48).

Now in the observational regime, $Y \perp\!\!\!\perp (X_0, H) \mid (Z, X_1)$. By (55), this must hold in all regimes. This gives

$$Y \perp\!\!\!\perp (F_0, F_1, X_0, H) \mid (Z, X_1). \tag{57}$$

Properties (54) and (57) are together equivalent to

$$Y \perp\!\!\!\perp (F_0, F_1, X_0, X_1^*, H) \mid (Z, X_1). \tag{58}$$

Combining (58) with (52) now yields (45).

$$\square$$

Augmented DAG Finally, having derived Figure 14 from assumptions (31)–(34), we can eliminate X_0^* and X_1^* from it. The relationships between the domain and regime variables are then represented by the augmented DAG of Figure 13, which can now be used to express and manipulate causal properties of the system, without further explicit consideration of the ITT variables—such consideration only being required in making the argument to justify this use.

10.2 General DAG

The case of a general DAG follows by extension of the arguments of §10.1 above. Consider a set of domain variables, with observational independencies represented by a DAG $\mathcal{D}$. We consider the variables in some total ordering consistent with the partial order of the DAG.

Some of the variables, say (in order) $(X_i : i = 1, \ldots, k)$, will be potential targets for intervention, with associated intention-to-treat variables (X_i^*) and intervention indicator variables (F_i). Let V_i denote the set of all the domain variables coming between X_{i-1} and X_i in the order. We thus have an ordered list $L = (V_1, X_1, \ldots, V_k, X_k, V_{k+1})$ of domain variables, some of which are possible targets for intervention.

Let pre_i denote the set of all predecessors of X_i in L, including X_i, and suc_i the set of all successors of X_i, excluding X_i. By pre_i^* we understand the set where all action variables in pre_i are replaced by their associated intention-to-treat variables, and similarly for suc_i^*. Also $F_{i:j}$ will denote $(F_i, \ldots, F_j)$, and similarly for other variables.

Generalising (31) with (32), or (33) with (34), and with similar motivation, we introduce the following assumptions (noting that B_i expresses a strong ignorability property for the effects of all the variables $(X_1, \ldots, X_i)$ on later variables—which would need correspondingly strong justification in any specific application):

$$A_i \quad : \qquad \mathrm{pre}_i^* \perp\!\!\!\perp F_{i:k} \mid F_{1:i-1} \tag{59}$$

$$B_i \quad : \qquad \mathrm{suc}_i^* \perp\!\!\!\perp (F_{1:i}, X_{1:i}^*) \mid (F_{i+1:k}, \mathrm{pre}_i). \tag{60}$$

Taking account of the fact that X_i is determined by (F_i, X_i^*), these are equivalent to:

$$A_i' \quad : \qquad (V_{1:i}, X_{1:i}^*, X_{1:i-1}) \perp\!\!\!\perp F_{i:k} \mid F_{1:i-1} \tag{61}$$

$$B_i' \quad : \qquad (V_{i+1:k}, X_{i+1:k}^*, X_{i+1:k}) \perp\!\!\!\perp (F_{1:i}, X_{1:i}^*) \mid (F_{i+1:k}, V_{1:i}, X_{1:i}). \tag{62}$$

Theorem 1 *Suppose the observational conditional independencies are represented by a DAG $\mathcal{D}$, and that assumptions A_i and B_i $(i = 1, \ldots, k)$ hold. Then the extended conditional independencies between domain variables, intention-to-treat variables, and regime variables (conditional on the regime variables) are represented by the ITT DAG $\mathcal{D}^*$, constructed by modifying $\mathcal{D}$ as follows:*

- *Each action variable X_i is replaced by the trio of variables F_i, X_i^* and X_i, with arrows from F_i and X_i^* to X_i. It is assumed that (30) holds.*

- *F_i is a founder node.*

- *X_i^* inherits all the original incoming arrows of X_i.*

- *X_i loses its original incoming arrows, but retains its original outgoing arrows.*

30

Proof. See Appendix A. □

Finally, on eliminating the intention-to-treat nodes (X_i^*) from the ITT DAG, the relationships between the domain variables and regime variables are represented by the augmented DAG $\mathcal{D}^\dagger$, constructed from $\mathcal{D}$ by adding, for each X_i, F_i as a founder node, with an arrow from F_i to X_i. As described in §2, such an augmented DAG is all we need to represent and manipulate causal properties. The above argument shows what needs to be assumed—and, more important, justified—to validate its use.

11 Comparison with other approaches

In this section we explore some of the similarities and differences between the decision-theoretic approach to statistical causality, considered above, and other currently popular approaches.

11.1 Potential outcomes

In the potential outcome (PO) formulation of statistical causality (Rubin 1974; Rubin 1978), the conception is that (for a generic individual) there exist, simultaneously and before the application of any treatment, two variables, $Y(0)$ and $Y(1)$: $Y(t)$ represents the individuals's *potential response* to the (actual or hypothetical) application of treatment t. If treatment 1 (resp., 0) is in fact applied, the corresonding potential outcome $Y(1)$ (resp., $Y(0)$) will be uncovered and so rendered actual, the observed response then being $Y = Y(1)$ (resp., $Y = Y(0)$); however the alternative, counterfactual, potential outcome $Y(0)$ (resp., $Y(1)$) will remain forever unobserved—a feature which Holland (1986) has termed the *fundamental problem of causal inference*, although it is not truly fundamental, but rather an artefact of the unnecessarily complicated PO approach.

The pair $(Y(1), Y(0))$ is supposed to have (jointly with the other variables in the problem) a bivariate distribution, common for all individuals—this might be regarded as generated from an assumption of exchangeability of the pairs $(Y_i(1), Y_i(0))$ across all individuals $i \in \mathcal{I}$. The marginal distribution of $Y(t)$ can be identified with our hypothetical distribution P_t for the (single) response variable Y under hypothesised application of treatment t, and is thus estimable from suitable experimental data. However, on account of the fundamental problem of causal inference no empirical information is obtainable about the dependence between $Y(0)$ and $Y(1)$, which can never be simultaneously observed.

Causal effect If I (individual 0) consider taking treatment 1 [resp., 0], I would then be looking forward to obtaining response $Y_0(1)$ [resp., $Y_0(0)$]. Causal interest, and inference, will thus centre on a suitable comparison between the two potential responses. The PO approach typically focuses on the "individual causal effect", ICE $:= Y(1) - Y(0)$. However, again on account of the fundamental problem of causal inference, ICE is never directly observable, and even its distribution can not be estimated from data except by making arbitrary and untestable assumptions (*e.g.*, that $Y(1)$ and $Y(0)$ are independent, or alternatively—"treatment-unit additivity, TUA"—that they differ by a non-random constant). For this reason attention is typically diverted to the *average causal effect*, ACE $:= \mathrm{E}(\mathrm{ICE})$. Since this can be re-expressed as $\mathrm{E}\{Y(1)\} - \mathrm{E}\{Y(0)\}$, and the individual expectations are estimable, so is ACE: indeed, although based on a different interpretation and expressed in different notation, it is essentially the same as our own definition (7) of ACE, which was introduced as one form of comparison between the two *distributions*, P_1 and P_0, for the single

31

response Y—rather than, as in the PO approach, an estimable distributional feature of the non-estimable comparison ICE between the two *variables* $Y(1)$ and $Y(0)$.

Consistency   In the PO approach, *consistency* refers to the property

$$Y = Y(T), \tag{63}$$

requiring that the response Y should be obtainable by revealing the potential response corresponding to the received treatment T. We can distinguish two aspects to this:

(i). When considered only in the context of an interventional regime $F_T = t$, (63) can be regarded as essentially a book-keeping device, since $Y(t)$ is *defined* as what would be observed if treatment t were applied.

(ii). But when it is understood as applying also in the observational regime, (63) has more bite, requiring that an individual's response to received treatment T should not depend on whether that treatment was applied by a (real or hypothetical) extraneous intervention, or, in the observational setting, by some unknown internal process. It is thus a not entirely trivial modularity assumption, forming the essential link between the observational and interventional regimes.

A parallel to aspect (i) in DT is the *temporal coherence* assumption appearing in footnote 8: this requires that uncertainty about the outcome Y, after it is known that treatment t has been applied, should be the same as the initial uncertainty about Y, on the hypothesis that treatment t will be applied. While not entirely vacuous, this too could be considered as little more than book-keeping.

More closely aligned with aspect (ii) is the distributional consistency property expressed in (11), which says that, for purposes of assessing the uncertainty about the response to a treatment t, the *only* difference between the interventional and the observational regime is that, in the latter, we have the additional information that the individual had been fingered to receive t. Again this has some empirical bite, and can be regarded as a not entirely trivial condition linking the observational and interventional regimes in the DT approach.

Treatment assignment and application   We have emphasised the distinction between the stochastic treatment assignment variable T^* and the non-stochastic treatment application indicator $\check{T}$. This is not explicitly done in the PO approach, but appears implicitly, since for any data individual, with fingered (and thus also actual) treatment T^* (typically just denoted by T in PO), we can distinguish between the actual response $Y = Y(T)$ in the observational regime, and the potential responses $Y(1)$ and $Y(0)$, relevant to the two interventional regimes.

We can make the following correspondences:

Ignorability   The PO expressions in (iv) and (v) of Table 1 have both been used to express ignorability in the PO framework, (iv) evidently being weaker than (v). The *weak ignorability* condition (iv) corresponds directly to the DT condition (9) for ignorability. However, the *strong ignorability* condition (v) has no DT parallel, since nothing in DT corresponds to a joint distribution of $(Y(0), Y(1))$. For applications weak ignorability (iv), which does have a DT interpretation, suffices. Similar remarks apply to the (weak and strong) *conditional ignorability* expressions in (vi) and (vii).

	PO	DT
(i)	Distribution of $Y(t)$	Distribution of Y given $\check{T} = t$
(ii)	Joint distribution of $(Y(0), Y(1))$	no parallel
(iii)	Distribution of Y given $T = t$	Distribution of Y given $T^* = t, \check{T} = t$
(iv)	$Y(t) \perp\!\!\!\perp T \quad (t = 0, 1)$	$Y \perp\!\!\!\perp T^* \mid \check{T}$
(v)	$(Y(0), Y(1)) \perp\!\!\!\perp T$	no parallel
(vi)	$Y(t) \perp\!\!\!\perp T \mid X \quad (t = 0, 1)$	$Y \perp\!\!\!\perp T^* \mid (X, \check{T})$
(vii)	$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$	no parallel

Table 1: Comparison of PO and DT approaches

SUTVA and SUTDA It is common in PO to impose the *Stable Unit-Treatment Value Assumption (SUTVA)* (**?**; **?**). This requires that, for any individual i, the potential response $Y_i(t)$ to application of treatment t to that individual should be unaffected by the treatments applied to other individuals. Indeed, without such an assumption the notation $Y_i(t)$ becomes meaningless, since the very concept intended by it is denied.

Our variant of SUTVA is the *Stable Unit-Treatment Distribution Assumption (SUTDA)*, as described in Condition 1. (Note that, unlike for SUTVA, even when this assumption fails it does not degenerate into meaninglessness, since the terms in it have interpretations independent of its truth). On making the further assumption, implicit in the PO approach, that, not just the set of values, but also the joint distribution, of the collection $\{Y_i(t) : i \in \mathcal{I}, t \in \mathcal{T}\}$ is unaffected by the application of treatments, it is easily seen that SUTVA implies SUTDA, so that our condition is weaker—and is sufficient for causal inference.

11.2   Pearlian DAGs

Judea Pearl has popularised graphical representations of causal systems based on DAGs. In §1.3 of Pearl (2009) he describes what he terms a "Causal Bayesian Network" (CBN), which we shall call a "Pearlian DAG".[16] This is intended to represent both the conditional independencies between variables in observational circumstances, and how their joint distributions changes when interventions are made on some or all of the variables: specifically, for any node not directly intervened on, its conditional distribution given its parents is supposed the same, no matter what other interventions are made.[17] The semantics of a Pearlian DAG representation is in fact identical with that, based entirely on d-separation, of the fully augmented observational DAG, in which every observable domain variable is accompanied by a regime indicator—thus allowing for the possibility of intervention on every such variable. However, although Pearl has occasionally included these regime indicators explicitly, as do we, for the most part he uses a representation where they are

---

[16]We avoid the term "causal DAG", which has been used with a variety of different interpretations (Dawid 2010).

[17]In the greater part of his causal writings, Pearl uses a different construction, in which all stochasticity is confined to unobservable "error variables", with domain variables related to these, and to each other, by deterministic functional relationships—he misleadingly terms this deterministic structure a "probabilistic causal model" (PCM). It is easy to show (Dawid 2002) that there is a many-one correspondence: any PCM implies a CBN structure for its domain variables, while any CBN can be derived from a, typically non-unique, PCM. Since the additional, unidentifiable, structure embodied in a PCM has no consequences for its use for decision-theoretic purposes, we do not consider these further here.

left implicit and omitted from the graph. A Pearlian DAG then looks, confusingly, exactly like the observational DAG, with its conditional independendies, but is intended to represent additional causal properties: properties that are *explicitly* represented by the corresponding augmented DAG.

Since such a Pearlian DAG is just an alternative representation of a particular kind of augmented DAG, its appropriateness must once again depend on the acceptability of the strong assumptions, described in § 10.2, needed to justify augmentation of an observational DAG.

11.3 SWIGs

Richardson and Robins (2013a); Richardson and Robins (2013b) introduce a different graphical representation of causal problems, the SWIG (single world intervention graph). A salient feature of this approach is "node-splitting", whereby a variable is represented twice: once as it appears naturally, and again as it responds to as intervention. Although the details of their representation and ours differ, they are based on similar considerations. Here we consider some of the parallels and differences between the two approaches.

Figure 3 of Richardson and Robins (2013a) (a single world intervention template, SWIT) is reproduced here as Figure 15, with notation changed so as more closely to match our own. Note the splitting of the treatment node T. As we shall see, this graph encodes ignorability of the treatment assignment, and can thus be compared with our own representations of ignorability.
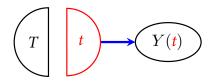

Figure 15: Simple SWIG template, expressing PO (weak) ignorability

In Figure 15, T denotes the treatment applied in the observational regime: it thus corresponds to our "intention-to-treat" variable T^*. The node labelled t represents an intervention to set the treatment to t: it therefore corresponds to $\check{T} = t$ in our development. The variable $Y(t)$, the "potential response" to the intervention at t, has no direct analogue in our approach, but that is inessential, since only its distribution is relevant; and that corresponds to our distribution P_t of Y in response to the intervention $\check{T} = t$.

Applying the standard d-separation semantics to Figure 15 (ignoring the unconventional shapes of some of the nodes), the disconnect between T and t represents their independence. This corresponds to our equation (8), encapsulating the covariate nature of T^*. Further, by the lack of an arrow from T to $Y(t)$, the graph encodes $Y(t) \perp\!\!\!\perp T$, which is to say that the distribution of $Y(t)$—the outcome consequent on a (real or hypothesised) intervention at t—is regarded as independent of the intention-to-treat variable (and this property should hold for all t). In our notation, this becomes $Y \perp\!\!\!\perp T^* \mid \check{T}$, as expressed in our equation (9), and represents ignorability of the treatment assignment. As described in § 7.1, in our treatment this can be represented by the DAG of Figure 7—which is therefore our translation of the SWIT of Figure 15, conveying essentially the same information in a different form.

Note that, in the approach of Richardson and Robins (2013a), in order fully to capitalise on the ignorability preperty represented by Figure 15, additional external use must be made of the

assumption of (functional) consistency: $T = t$ implies $Y(t) = Y$. For example, in this approach the average causal effect ACE is defined as $\mathrm{E}\{Y(0) - Y(1)\}$. Now by ignorability, as represented in the SWIT of Figure 15, $Y(t) \perp\!\!\!\perp T$, whence $\mathrm{E}\{Y(t)\} = \mathrm{E}\{Y(t) \mid T = t\}$. But we then need to make further use of functional consistency to replace this by $\mathrm{E}\{Y \mid T = t\}$, so obtaining ACE $= \mathrm{E}\{Y \mid T = 1\} - \mathrm{E}\{Y \mid T = 0\}$.

Our analogue of functional consistency is distributional consistency (Definition 2): $Y \mid (T = t, F_T = \emptyset) \approx Y \mid (T^* = t, F_T = t)$. However, this property has already been used in justifying the representation by means of Figure 7. Once that graph is constructed, distributional consistency does not require further explicit attention since, as shown in Remark 1, it is already represented in Figure 5, and thus in Figure 7. And then Figure 8 can be used directly to represent and manipulate the fundamental DT representation of ignorability, as expressed by (18). Thus we define ACE $= \mathrm{E}(Y \mid F_T = 1) - \mathrm{E}(Y \mid F_T = 0)$. With ignorability expressed as $Y \perp\!\!\!\perp F_T \mid T$, as encoded in Figure 7, we immediately have $\mathrm{E}(Y \mid F_T = t) = \mathrm{E}(Y \mid T = t, F_T = t) = \mathrm{E}(Y \mid T = t)$, and thus ACE $= \mathrm{E}(Y \mid T = 1) - \mathrm{E}(Y \mid T = 0)$.

A further conceptual advantage of our approach is that is unnecessary to consider (even one-at-a-time) the distinct potential responses[18] $Y(t)$: we have a single response variable Y, but with a distribution that may be regime-dependent.

12   A comparative study: g-computation

In this Section we compare, contrast, and finally unify, the various approaches to causal modelling and inference, in the context of the specific example of § 10.1. We suppose we have observational data, and wish to identify the distribution of Y under interventions at X_0 and X_1. Purely for notational simplicity, we assume all variables are discrete

12.1   Pearl's *do*-calculus

The *do*-calculus (Pearl 2009, § 3.4) is a methodology for discovering when and how, for a problem represented by a specified Pearlian DAG, it is possible to use observational information to identify an interventional distribution. Notation such as $p(x \mid y, \widehat{z})$ refers to the distribution of X given the observation $Y = y$, when Z is set by intervention to z. Pearl gives 3 rules, based on interrogation of the DAG, that allow transformation of such expressions. If by successive application of these rules we can re-express our desired interventional target by a hatless expression, we are done.

In this notation, we would like to identify $p(y \mid \widehat{x}_0, \widehat{x}_1)$. We can write

$$p(y \mid \widehat{x}_0, \widehat{x}_1) = \sum_z p(y \mid \widehat{x}_0, \widehat{x}_1, z) \times p(z \mid \widehat{x}_0, \widehat{x}_1). \tag{64}$$

According to Pearl's Rule 2, we have

$$p(y \mid \widehat{x}_0, \widehat{x}_1, z) = p(y \mid x_0, x_1, z) \tag{65}$$

because Y is d-separated from (X_0, X_1) by Z in the DAG of Figure 12 modified by deleting the arrows out of X_0 and X_1.

---

[18]unhelpfully described as "counterfactuals" by Richardson and Robins (2013a)

Next, again by Rule 2, we can show

$$p(z \mid \widehat{x}_0, \widehat{x}_1) = p(z \mid x_0, \widehat{x}_1) \tag{66}$$

by seeing that Z is d-separated from X_0 by X_1 in the DAG modified by the deleting arrows into X_1 and out of X_0.

Finally, by Rule 3, we confirm

$$p(z \mid x_0, \widehat{x}_1) = p(z \mid x_0) \tag{67}$$

because Z is d-separated from X_1 by X_0 in the DAG with arrows into X_1 removed. So on combining (66) and (67) we have shown

$$p(z \mid \widehat{x}_0, \widehat{x}_1) = p(z \mid x_0). \tag{68}$$

Inserting (65) and (68) into (64), we conclude

$$p(y \mid \widehat{x}_0, \widehat{x}_1) = \sum_z p(y \mid x_1, z) \times p(z \mid x_0), \tag{69}$$

showing that the desired interventional distribution can be constructed from ingredients identifiable in the observational regime. Equation (69) is (a simple case of) the *g-computation* formula of Robins (1986).

12.2 DT approach

As described in Dawid (2015), the DT approach supplies a more straightforward way of justifying and implementing *do*-calculus, using the augmented DAG. In our problem this is Figure 13, and what we want is $p(Y = y \mid F_0 = x_0, F_1 = x_1)$.

Noting $F_0 = x_0 \Rightarrow X_0 = x_0$ *etc.*, in general we have:

$$
\begin{aligned}
p(Y = y \mid F_0 = x_0, F_1 = x_1) \;=\; & \sum_z p(Y = y \mid X_0 = x_0, X_1 = x_1, Z = z, F_0 = x_0, F_1 = x_1) \\
& \times p(Z = z \mid X_0 = x_0, F_0 = x_0, F_1 = x_1).
\end{aligned} \tag{70}
$$

Applying d-separation to Figure 13, we can infer the following conditional independencies:

$$
\begin{aligned}
Y \;&\perp\!\!\!\perp\; (F_0, X_0, F_1) \mid (Z, X_1) \tag{71} \\
Z \;&\perp\!\!\!\perp\; (F_0, F_1) \mid X_0. \tag{72}
\end{aligned}
$$

Using these in (70) we obtain

$$
\begin{aligned}
p(Y = y \mid F_0 = x_0, F_1 = x_1) \;=\; & \sum_z p(Y = y \mid X_1 = x_1, Z = z, F_0 = \emptyset, F_1 = \emptyset) \\
& \times p(Z = z \mid X_0 = x_0, F_0 = \emptyset, F_1 = \emptyset), \tag{73}
\end{aligned}
$$

which is (69), re-expressed in DT notation.

12.3 PO approach

The Pearlian/DT approach makes no use of potential outcomes. By contrast, these are fundamental to the original approach of Robins, where the conditions supporting g-computation are:

$$Y(x_0, x_1) \quad \perp\!\!\!\perp \quad X_1 \mid (Z, X_0 = x_0) \tag{74}$$

$$Z(x_0) \quad \perp\!\!\!\perp \quad X_0. \tag{75}$$

In his Example 11.3.3, Pearl (2009), basing his argument on his "twin-network" construction, claims that (74) can not be derived from a PO interpretation of Figure 12. However, Richardson and Robins (2013a) refute this by constructing the SWIT version of Figure 12, as in Figure 16.
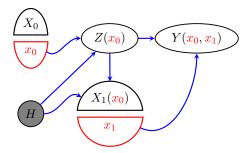

Figure 16: SWIT

This DAG encodes the property

$$Y(x_0, x_1) \perp\!\!\!\perp X_1(x_0) \mid (Z(x_0), X_0)$$

whence

$$Y(x_0, x_1) \perp\!\!\!\perp X_1(x_0) \mid (Z(x_0), X_0 = x_0). \tag{76}$$

They then apply functional consistency, $X_0 = x_0 \Rightarrow Z(x_0) = Z, X_1(x_0) = X_1$, to deduce (74). As for (75), this is directly encoded in Figure 16.

12.4 Unification

We can use the DT approach to relate all the approaches above.

12.4.1 DT for SWIG/PO

Figure 14, using explicit "intention to treat" variables and regime indicators, is the DT reinterpretation of the SWIT of Figure 16.

From Figure 14 (noting that the dotted arrow from X_1^* to X_1 disappears when $F_1 \neq \emptyset$) we can read off

$$Y \perp\!\!\!\perp X_1^* \mid Z, X_0, F_0, F_1 = x_1,$$

so that

$$Y \perp\!\!\!\perp X_1^* \mid Z, X_0 = x_0, F_0 = \emptyset, F_1 = x_1, \tag{77}$$

which is the DT paraphrase of (74). Similarly the DT paraphrase of (75),

$$Z \perp\!\!\!\perp X_0^* \mid F_0 = x_0, \tag{78}$$

is likewise encoded in Figure 14. (In particular, both these properties are consequences of our assumptions (31)–(34), together with (30).)

12.4.2 Consistency?

Note that the derivations in § 12.4.1 above do not require further explicit application of (functional or distributional) consistency conditions. We could have complicated the analysis by mimicking more closely that of § 12.3. The DT paraphrase of (76), which can be read off Figure 14, is

$$Y \perp\!\!\!\perp X_1^* \mid Z, X_0^*, F_0 = x_0, F_1 = x_1.$$

On restricting to $X_0^* = x_0$ and applying the distributional consistency condition, we obtain the DT paraphrase of (74):

$$Y \perp\!\!\!\perp X_1^* \mid Z, X_0 = x_0, F_0 = \emptyset, F_1 = x_1.$$

But note that the required distributional consistency property can be expressed as

$$Y \perp\!\!\!\perp (X_1^*, F_0) \mid (Z, X_0, F_1 = x_1),$$

and this is already directly encoded in Figure 14. That being the case, we can leave it implicit and shortcut the analysis, as in § 12.4.1

12.4.3 DT for Pearl

We have shown that, if we can justify the DT ITT representation of Figure 14, we can derive (74) and (75), the conditions used to derive the g-computation formula (69) in the PO approach. However, the same end-point can be reached much more directly. Extracting from Figure 14 the conditional independencies between just the observable variables and the intervention indicators (*i.e.*, eliminating X_0^* and X_1^*), we recover Figure 13, the DT version of the Pearlian DAG Figure 12. From this, as shown in § 12.2, (69) can readily be deduced directly, without any need to complicate the analysis by consideration of potential outcomes. As described in § 10.1.1, consideration of intention-to-treat variables is needed to justify the appropriateness of the augmented/Pearlian DAG of Figure 13; but once that has been done, for further analysis we can simply forget about the ITT variables X_0^* and X_1^*.

Dawid and Didelez (2010), § 10.1.1, show how the PO conditions typically imposed to justify more general forms of g-computation imply the much simpler DT conditions supporting more straightforward justification. The DT approach can, moreover, be straightforwardly extended to allow sequentially dependent randomised interventions, which can introduce considerable additional complications for the PO approach.

13 Discussion

In this paper we have developed a clear formalism for problems of statistical causality, based on the idea that I want to use external data to assist me in making a decision. We have shown

how this serves as a firm theoretical foundation for methods framed within the DT approach, enabling transfer of probabilistic information from an observational to an interventional setting. We have emphasised, in particular, just what considerations are involved—and so what needs to be argued for—when we invoke enabling assumptions such as ignorability. In the course of the development we have introduced DT analogues of concepts arising in other causal frameworks, including consistency and the stable unit-treatment value assumption, and clarified the similarities and differences between the different approaches.

General though our analysis has been, it could be generalised still further. For example, our exchangeability assumptions treat all individuals on a par. But we could consider more complex versions of exchangeability, such as are relevant in experimental designs where we distinguish various factors which may be crossed or nested (Dawid (1988), Dawid (2000) § 10.1); or conducted more detailed modelling of non-exchangeable data. Our analysis of DAGs in this article has been restricted to non-randomised point interventions, taking no account of information previously learned. Further extension would be needed to fully justify, *e.g.*, DT models for dynamic regimes (Dawid and Didelez 2010).

References

Berzuini, C., Dawid, A. P., and Bernardinelli, L. (ed.) (2012a). *Causality: Statistical Perspectives and Applications*. John Wiley & Sons, Ltd, Chichester, UK.

Berzuini, C., Dawid, A. P., and Didelez, V. (2012b). Assessing dynamic treatment strategies. In (Berzuini *et al.* 2012a), chapter 8, pp. 85–10.

Bühlmann, P. (2018). Invariance, causality and robustness. arXiv:1812.08233.

Constantinou, P. and Dawid, A. P. (2017). Extended conditional independence and applications in causal inference. *Annals of Statistics*, **45**, 2618–53.

Dawid, A. P. (1979a). Conditional independence in statistical theory (with Discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 1–31.

Dawid, A. P. (1979b). Some misleading arguments involving conditional independence. *Journal of the Royal Statistical Society, Series B*, **41**, 249–52.

Dawid, A. P. (1980). Conditional independence for statistical operations. *Annals of Statistics*, **8**, 598–617.

Dawid, A. P. (1988). Symmetry models and hypotheses for structured data layouts (with Discussion). *Journal of the Royal Statistical Society, Series B*, **50**, 1–34.

Dawid, A. P. (2000). Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association*, **95**, 407–48.

Dawid, A. P. (2001). Some variations on variation independence. In *Artificial Intelligence and Statistics 2001*, (ed. T. Jaakkola and T. S. Richardson), pp. 187–91. Morgan Kaufmann Publishers, San Francisco, California.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**, 161–89. Corrigenda, *ibid.*, 437.

Dawid, A. P. (2003). Causal inference using influence diagrams: The problem of partial compliance (with Discussion). In *Highly Structured Stochastic Systems*, (ed. P. J. Green, N. L. Hjort, and S. Richardson), pp. 45–81. Oxford University Press.

Dawid, A. P. (2007a). Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality. In *Causality and Probability in the Sciences*, Texts in Philosophy, Vol. 5, (ed. F. Russo and J. Williamson), pp. 503–32. College Publications, London.

Dawid, A. P. (2007b). Fundamentals of statistical causality. Research Report 279, Department of Statistical Science, University College London. 94 pp.
`https://www.ucl.ac.uk/drupal/site_statistics/sites/statistics/files/migrated-files/rr279.pdf` .

Dawid, A. P. (2010). Beware of the DAG! In *Proceedings of the NIPS 2008 Workshop on Causality*, Journal of Machine Learning Research Workshop and Conference Proceedings, Vol. 6, (ed. I. Guyon, D. Janzing, and B. Schölkopf), pp. 59–86. `http://tinyurl.com/33va7tm` .

Dawid, A. P. (2012). The decision-theoretic approach to causal inference. In (Berzuini *et al.* 2012a), chapter 4, pp. 25–42.

Dawid, A. P. (2015). Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and its Application*, **2**, 273–303.
`DOI:10.1146/annurev-statistics-010814-020105`.

Dawid, A. P. and Constantinou, P. (2014). A formal treatment of sequential ignorability. *Statistics in Biosciences*, **6**, 166–88.

Dawid, A. P. and Didelez, V. (2008). Identifying optimal sequential decisions. In *Proceedings of the Twenty-Fourth Annual Conference on Uncertainty in Artificial Intelligence* (UAI-08), pp. 113–20. AUAI Press, Corvallis, Oregon.
`http://uai2008.cs.helsinki.fi/UAI_camera_ready/dawid.pdf`.

Dawid, A. P. and Didelez, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistical Surveys*, **4**, 184–231.

de Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, **7**, 1–68. English translation "Foresight: Its Logical Laws, Its Subjective Sources" by H. E. Kyburg, in Kyburg and Smokler (1964), 55–118.

de Finetti, B. (1938/1980). On the condition of partial exchangeability. In *Studies in Inductive Logic and Probability*, (ed. R. C. Jeffrey), pp. 193–205. University of California Press, Berkeley, Los Angeles, London.

de Finetti, B. (1975). *Theory of Probability (Volumes 1 and 2)*. John Wiley and Sons, New York. (Italian original Einaudi, 1970).

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

Didelez, V., Dawid, A. P., and Geneletti, S. G. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pp. 138–46. AUAI Press, Arlington, Virginia.

Dowe, P. (2000). *Physical Causation*. Cambridge University Press, Cambridge.

Geiger, D., Verma, T. S., and Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, **20**, 507–34.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society, Series B*, **69**, 199–215.

Geneletti, S. and Dawid, A. P. (2011). Defining and identifying the effect of treatment on the treated. In *Causality in the Sciences*, (ed. P. M. Illari, F. Russo, and J. Williamson), pp. 728–49. Oxford University Press.

Guo, H. and Dawid, A. P. (2010). Sufficient covariates and linear propensity analysis. *Journal of Machine Learning Research Workshop and Conference Proceedings*, **9**, 281–8. Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna, Sardinia, Italy, May 13–15, 2010, edited by Y. W. Teh and D. M. Titterington. `http://jmlr.csail.mit.edu/proceedings/papers/v9/guo10a/guo10a.pdf`.

Guo, H., Dawid, A. P., and Berzuini, G. M. (2016). Sufficient covariate, propensity variable and doubly robust estimation. In *Statistical Causal Inferences and Their Applications in Public Health Research*, (ed. H. He, P. Wu, and D.-G. Chen), pp. 49–89. Springer. `DOI:10.1007/978-3-319-41259-7_3`.

Hausman, D. (1998). *Causal Asymmetries*. Cambridge University Press, Cambridge.

Heckman, J. J. (1992). Randomization and social policy evaluation. In *Evaluating Welfare and Training Programs*, (ed. C. F. Manski and I. Garfinkel), chapter 5, p. 20123. Harvard University Press, Cambridge, MA.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, **17**, 360–72.

Holland, P. W. (1986). Statistics and causal inference (with Discussion). *Journal of the American Statistical Association*, **81**, 945–970.

Janzing, D. and Schölkopf, B. (2010). Distinguishing between cause and effect using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, **56**, 5168–94.

Kyburg, H. E. and Smokler, H. E. (ed.) (1964). *Studies in Subjective Probability*. John Wiley and Sons, New York.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491–505.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Analytical methods for social research, (Second edn). Cambridge University Press.

Pearl, J. (1988). *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann Publishers, San Mateo, California.

Pearl, J. (1993a). Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, pp. 391–401.

Pearl, J. (1993b). Comment: Graphical models, causality and intervention. *Statistical Science*, **8**, 266–9.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, (Second edn). Cambridge University Press, Cambridge.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, (ed. W. Burgard and D. Roth), pp. 247–54. AAAI Press, Menlo Park, CA. `http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3769/3864`.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why*. Basic Books, New York.

Price, H. (1991). Agency and probabilistic causality. *British Journal for the Philosophy of Science*, **42**, 157–76.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. MIT Press, Cambridge, Massachusetts.

Reichenbach, H. (1956). *The Direction of Time*. University of Los Angeles Press, Berkeley.

Richardson, T. S. and Robins, J. M. (2013a). Single world intervention graphs: A primer. Second UAI Workshop on Causal Structure Learning, Bellevue, Washington, July 15 2013.

Richardson, T. S. and Robins, J. M. (2013b). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report 128, Center for Statistics and Social Sciences, University of Washington.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–512.

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.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test— Comment. *Journal of the American Statistical Association*, **75**, (371), 591–3.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, **81**, (396), 961–2.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton.

Skyrms, B. (1987). Dynamic coherence and probability kinematics. *Philosophy of Science*, **54**, 1–20.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*, (Second edn). Springer-Verlag, New York.

Spohn, W. (2001). Bayesian nets are all there is to causal dependence. In *Stochastic Dependence and Causality*, (ed. M. C. Galavotti, P. Suppes, and D. Costantini), chapter 9, pp. 157–72. University of Chicago Press, Chicago.

Suppes, P. (1970). *A Probabilistic Theory of Causality*, Acta Philosophica Fennica, Vol. 24. North-Holland, Amsterdam.

Webb, R. (2020). Finding our place in the universe. "New Scientist" article, 15 February 2020.
`https://institutions.newscientist.com/article/mg24532690-700-your-decision-making-ability-is-a-superpower-`
.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford.

Woodward, J. (2016). Causation and manipulability. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.).
`https://plato.stanford.edu/entries/causation-mani/`.

A Proof of Theorem 1

As in Remark 5, and purely as an instrumental tool, we regard all the regime variables as stochastic and mutually independent:

$$\underset{i=1}{\overset{k}{\perp\!\!\!\perp}} F_i. \tag{79}$$

We shall show that $\mathcal{D}^*$ then represents the conditional independencies between all its variables. The desired result will then follow on conditioning on $F_{1:k}$.

For economy of notation, we write W_i for (V_i, X_i), $W_{a:b}$ for $(V_{a:b}, X_{a:b})$. and similarly W_i^*, $W_{a:b}^*$.

Lemma 5 *For each $r = 1, \ldots, k-1$,*

$$H_r : F_{r+1:k} \perp\!\!\!\perp F_{1:r} \mid W_{1:r}. \tag{80}$$

Proof. We show (80) by induction.

By (79), $F_1 \perp\!\!\!\perp F_{2:k}$, while by A_2' we have $W_1 \perp\!\!\!\perp F_{2:k} \mid F_1$. Together these yield $(F_1, W_1) \perp\!\!\!\perp F_{2:k}$, from which H_1 follows.

Suppose now H_r holds. From B_r' we have

$$W_{r+1} \perp\!\!\!\perp F_{1:r} \mid (F_{r+1:k}, W_{1:r}). \tag{81}$$

Together with H_r this gives

$$(F_{r+1:k}, W_{r+1}) \perp\!\!\!\perp F_{1:r} \mid W_{1:r} \tag{82}$$

whence

$$F_{r+1} \perp\!\!\!\perp F_{1:r} \mid (F_{r+2:k}, W_{1:r+1}). \tag{83}$$

Also, by A_{r+2}',

$$W_{1:r+1} \perp\!\!\!\perp F_{r+2:k} \mid F_{1:r+1}, \tag{84}$$

which together with $F_{1:r+1} \perp\!\!\!\perp F_{r+2:k}$, from (79), gives $(F_{1:r+1}, W_{1:r+1}) \perp\!\!\!\perp F_{r+2:k}$, from which we have

$$F_{r+2:k} \perp\!\!\!\perp F_{1:r+1} \mid W_{1:r+1}. \tag{85}$$

So H_{r+1} holds and the induction is established. $\qquad\square$

Lemma 6 *For each r:*

$$(V_{r+1}, X_{r+1}^*) \perp\!\!\!\perp (F_{1:k}, X_{1:r}^*) \mid (V_{1:r}, X_{1:r}). \tag{86}$$

Proof.

From B_r', we have

$$W_{r+1}^* \perp\!\!\!\perp F_{1:r} \mid (F_{r+1:k}, W_{1:r}). \tag{87}$$

Combining this with (80) gives

$$(F_{r+1:k}, W_{r+1}^*) \perp\!\!\!\perp F_{1:r} \mid W_{1:r},$$

whence

$$W_{r+1}^* \perp\!\!\!\perp F_{1:r} \mid W_{1:r}. \tag{88}$$

43

Also, from A'_{r+1},
$$W^*_{r+1} \perp\!\!\!\perp F_{r+1:k} \mid (F_{1:r}, W_{1:r}).$$

Together with (88) this gives
$$W^*_{r+1} \perp\!\!\!\perp F_{1:k} \mid W_{1:r}. \tag{89}$$

Also from B'_r we have
$$W^*_{r+1} \perp\!\!\!\perp X^*_{1:r} \mid (F_{1:k}, W_{1:r}). \tag{90}$$

Now combining (89) and (90) we obtain (86). $\qquad\square$

To complete the proof of Theorem 1, consider the sequence
$$L^* = (F_1, \ldots, F_k, V_1, X^*_1, X_1, \ldots, V_k, X^*_k, X_k, V_{k+1}).$$

which is consistent with the partial order of the ITT DAG $\mathcal{D}^*$. Each V_i may comprise a number of domain variables: we consider it as expanded into its constituent parts, respecting the partial order of $\mathcal{D}$, and thus of $\mathcal{D}^*$.

To establish Theorem 1, we show that each variable in L^* is independent of its predecessors in L^*, conditional on its parent variables in $\mathcal{D}^*$.

(i). For each F_i, this holds by (79).

(ii). For an intervention target X_i, its only parents in $\mathcal{D}^*$ are X^*_i and F_i. By (30), conditional on these X_i is fully determined, hence independent of anything.

(iii). Consider now a non-intervention domain variable, U say. Its parents in $\mathcal{D}^*$ are the same as its parents in $\mathcal{D}$. Now U is contained in V_r for some r. By (86) its conditional distribution, given all its predecessors in L^*, depends only on the preceding domain variables. In particular, this conditional distribution, being the same in all regimes, must agree with that in the observational regime, whose independencies are encoded in the initial DAG $\mathcal{D}$—and so depends only on the parents of U in $\mathcal{D}$, and hence in $\mathcal{D}^*$.

(iv). The remaining case, of an ITT variable X^*_i, follows similarly to (iii), on further noting that the parents of X^*_i in $\mathcal{D}^*$ are the same as the parents of X_i in $\mathcal{D}$, and X^*_i is identical to X_i in the observational setting.