The Logic of Counterfactuals and the Epistemology of Causal Inference

Hanti Lin

University of California, Davis

ika@ucdavis.edu

November 15, 2024

Abstract

The 2021 Nobel Prize in Economics recognized a theory of causal inference that warrants more attention from philosophers. To this end, I design a tutorial on that theory for philosophers and develop a dialectic that connects to a traditional debate in philosophy: the Lewis-Stalnaker debate on Conditional Excluded Middle (CEM). I first defend CEM, presenting a new Quine-Putnam indispensability argument based on the Nobel-winning application of the Rubin causal model (the potential outcome framework). Then, I switch sides to challenge this argument, introducing an updated version of the Rubin causal model that preserves the successful application while dispensing with CEM.

1 Introduction

This is an invitation to the *Rubin causal model* (Rubin 1974), also known as the *potential outcome framework*—a framework for causal inference that has been very influential in health and social sciences but is somewhat under-recognized in philosophy. To make more philosophers interested, I will explain how the Rubin causal model is related to some familiar ideas and issues in philosophy, such as causal Bayes nets, intertheory relations, the Quine-Putnam indispensability argument, the revisability of deductive logic, and the controversy over a logical principle called:

Conditional Excluded Middle

It is logically necessary that
- either B would be the case if A were the case,
- or B would *not* the be case if A were the case.

To make all this more fun, a dialectic will be developed. I will first give a new argument for Conditional Excluded Middle, using the Rubin causal model and its Nobel-Prize winning applications (section 3). But then, following the good cop/bad cop approach, I will challenge that argument with a new theory of causal inference.

Before all that, I will begin with a crash course on the Rubin causal model (next section). Since an important goal of this paper is to introduce the Rubin causal model to a broad philosophical audience, I have taken care to distill the essential ideas and present them accessibly in the main text, which should suffice for most philosophers. A rigorous presentation is provided in the appendices.

2 A Crash Course on the Rubin Causal Model: A Card Game

To have a vivid picture of how the Rubin causal model works, imagine that everyone in the population has such a card:

Card #1: What If You Took the Treatment?

Nature gives every individual a card of this form: the back is printed with '*if Take* = 1', and the face is printed with '*Cure* = 1' or '*Cure* = 0'.

The former case means that this person would be cured if they took the treatment; the latter means that this person would *not* be cured if they took the treatment. So this setting builds in Conditional Excluded Middle. Any card given to a person is face down initially and will be flipped to reveal the (medical) result only when the if-clause actually holds of that person. Similarly, there is also

Card #2: What If You Didn't Take the Treatment?

Nature also gives every individual a second card, whose face takes the same form '*Cure* = ...' but the back is printed with '*if Take* = 0' instead.

Each person's cards #1 and #2 define that person's *individual treatment effect* (ITE): the value of binary variable *Cure* on card #1 minus that on card #2. So there are three possible cases:

$$\text{ITE} \;=\; \begin{cases} 1 & (=1-0) & \text{i.e. improvement,} \\ 0 & (=1-1 \text{ or } 0-0) & \text{i.e. no difference,} \\ -1 & (=0-1) & \text{i.e. deterioration.} \end{cases}$$

The *average treatment effect* (ATE) in a population is defined as the average of the individual treatment effects of all individuals in the population.

A bit of algebra shows that the ATE is equal to the difference between two proportions:

$$\begin{aligned} \text{ATE} \;=\;& \text{(i) the proportion of `}\textit{Cure} = 1\text{' cards among all cards of the kind \#1} \\ & - \text{(ii) the proportion of `}\textit{Cure} = 1\text{' cards among all cards of the kind \#2.} \end{aligned}$$

Term (i) can be estimated by randomly choosing a group of people in the population, *forcing* them to flip their first cards, and registering the proportion of the results that have '$\textit{Cure} = 1$'. Term (ii) can estimated similarly. This estimation procedure is the idea behind randomized controlled trials (RCT). But the problem is that RCT is often ethically impermissible.

Fortunately, there is a Nobel-winning solution. Randomly select people from the population and flip a coin to assign them to the treatment or the control group. Anyone in the treatment group is offered the treatment for free—they decide whether to take it. This creates a new kind of card:

> **Card #3: What If You Were Assigned to the Treatment Group?**
>
> Nature also gives every individual a card of this form: the back is printed with '*if Assign* $= 1$' (where 1 means the treatment group), and the face is printed with '$\textit{Take} = 1$' or '$\textit{Take} = 0$'.

This determines whether the individual would, or would not, take the treatment under assignment to the treatment group. Similarly:

> **Card #4: What If You Were Assigned to the Control Group?**
>
> Nature gives every individual a card of this form: the back is printed with

'*if Assign* = 0' (where 0 means the control group), and the face is printed with '*Take* = 1' or '*Take* = 0'.

While seemingly unnecessary in the case of drug test, this fourth card is crucial in Angrist's (1990) classic study on the Vietnam War. There,"assignment" is the draft lottery, "treatment" is military service, and the "medical result" is lifetime earnings. The fourth card is needed to define volunteers.

Now, let's distinguish four subpopulations, depending on whether one would (or would not) take the treatment under assignment to the treatment group (or the control group):

1. *Compilers*: those who would take the treatment if they were assigned to the treatment group, and would not if they were assigned to the control group—namely, those whose card #3 and card #4 are printed with '*Take* = 1' and '*Take* = 0' respectively.

2. *Defiers*: those who would do the opposite to what compliers would do.

3. *Always-Takers*: those who would take the treatment anyway.

4. *Never-Takers*: those who would not take the treatment anyway.

By Conditional Excluded Middle, those four subpopulations exhaust the entire population.

An important result in econometrics implies that, in this card game scenario and in more general settings, if there are *no defiers*, which seems to be plausible to assume here, then we can "nicely" estimate a local average treatment effect (LATE), defined as the average of the individual treatment effects of just the *compliers* in the population. This classic result can be informally stated as follows, with a rigorous presentation provided in Appendix A and a proof in Appendix B:

> **Theorem 1 (Imbens & Angrist 1994, Informal Version).** *Under the eight assumptions are made precise in Appendix A (which are also instantiated by the card game informally designed above), the* LATE *in the compliers can be expressed solely by probabilities over the three observable variables—Assign, Take, and Cure—without counterfactuals. Specifically, the* LATE *can be expressed as follows:*
>
> $$\text{LATE} \;=\; \frac{\Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=0)}{\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=0)}\,.$$

An explanation of the probability function Pr is in order. The first conditional probability on the right-hand side, $\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 1)$, is defined standardly as a ratio:

$$\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 1) \;\; = \;\; \frac{\Pr(\mathit{Cure} = 1 \wedge \mathit{Assign} = 1)}{\Pr(\mathit{Assign} = 1)} \,,$$

which is the probability of drawing an individual from the population who ends up being assigned to the treatment group ($\mathit{Assign} = 1$) and then getting cured ($\mathit{Cure} = 1$), divided by the probability of drawing one who ends up being assigned to the treatment group. Thanks to elementary statistics, this probability has a nice[1] estimator that almost all frequentist statisticians agree on: just estimate it by the proportion of the cured in the treatment group. The other conditional probabilities on the right are also defined standardly and can be similarly estimated by the observed proportions. This procedure for estimating the right side, and thus the left side, is called *instrumental variable estimation*, with *Assign* being the *instrumental variable*.

The above theorem does something important. Recall that the LATE is defined in terms of the contents of many cards that cannot be flipped to reveal their faces at the same time. Indeed, within the entire population, all cards of types #3 and #4 are used to define the subpopulation of compliers. Then, within this subpopulation, all cards of types #1 and #2 are used to define the local average treatment effect, which is the proportion of '$\mathit{Cure} = 1$' among cards of type #1 (in this subpopulation) minus that among cards of type #2. But we cannot flip both card #1 and card #2 of any particular person's—it is impossible for a single person to take the treatment and to not take it. Similarly, we cannot flip both card #3 and card #4 of any particular person's. Even if God, or Nature, has the privilege to peek at the contents of all cards, we don't. Fortunately, to estimate the LATE, it suffices to estimate the probabilities on the right-hand side of the equation in Theorem 1. That does not require us to flip all those cards, but only require us to flip a card only when the if-clause printed on it actually holds—in accordance with the rule of the card game that Nature imposes on us as human beings.

Upshot: In the present scenario, the causal effect defined as the LATE (on the left side) can be *identified* in terms of some probabilities (on the right) that, in turn, can be nicely estimated with a simple statistical procedure—despite the fact that causation

[1]Namely, unbiased and (statistically) consistent.

cannot be defined in purely statistical terms. So, this theorem is known as an *identification* result. This result and its applications underly one half of the 2021 Nobel Prize in Economics, awarded to Joshua D. Angrist and Guido W. Imbens.

The proof of Theorem 1 is available in research articles, but may be difficult to follow for most philosophers. Even a textbook presentation of the proof is often too terse for most philosophers and, worse, typically occurs only after two hundred pages of discussion of the Rubin causal model, e.g. Hernán & Robins (2020: Technical Point 16.6), with the presuppositions of the proof somewhat scattered in preceding chapters. So, to make the materials self-contained and accessible for a wider community of philosophers, an alternative proof is designed in Appendix A, with minimal prerequisites for philosophers (it suffices to have some familiarity with elementary probability theory). This finishes the first task of this paper—a crash course on the Rubin causal model and the identification result of the LATE for philosophers.

All proofs of Theorem 1 in the existing literature presupposes that the population is exhausted by the four subpopulations defined above, and that in turn presupposes Conditional Excluded Middle, which brings us to:

3 The Lewis-Stalnaker Debate on CEM

Conditional Excluded Middle (CEM) sparks debate in philosophy of language—embraced by Stalnaker (1968) and rejected by Lewis (1973).

To quickly review an influential argument against CEM (Lewis 1973, Hájek MS), consider the following pair of sentences:

(A) If i had taken the treatment, i would have been cured.
(B) If i had taken the treatment, i would not have been cured.

CEM says that $(A) \vee (B)$ is true in every possible world. To find a counterexample, think about an indeterministic world in which the following holds:

(C) If i had taken the treatment, i would have had a (probabilistic) chance p to be cured and a chance $1 - p$ to be not cured, where p lies strictly between 0 and 1.

Then argue as follows that the truth of (C) implies the falsity of both (A) and (B):

Indeterminist Argument Against CEM

1. By (C), if i had taken the treatment, i would have had a nonzero chance to be cured and a nonzero chance to be not cured.

2. So, by (1), if i had taken the treatment, i could have been cured and could have been not cured.

3. Then (A) is false, for it contradicts (2).

4. Similarly, (B) is also false, for it contradicts (2).

5. So, by (3) and (4), disjunction $(A) \vee (B)$ is false.

Hence a counterexample to CEM in such an indeterministic world—or so the Lewisian concludes.

The above is round one. The next round will feature responses from defenders of CEM, such as Williams (2010); for a survey of this debate see Mandelkern (2022, sec. 17.3.4). Here is the thing: defenders of CEM should also explore a new argument in their favor.

Indispensability Argument For CEM

> CEM is assumed, and seems to be indispensable, in our best theory of causal inference in health and social sciences—the theory that led to one half of the 2021 Nobel Prize in Economics. Indeed, the assumption of this logical principle has long been made since the early days of this theoretical framework (Rubin 1974). Moreover, if we take a close look at the proof strategy for Theorem 1 as presented in Appendix B, it does seem that the assumption of CEM is essential. So, it seems that we should accept CEM.

This argument is patterned after the Quine-Putnam indispensability argument for the existence of certain mathematical objects: the mathematical objects that are indispensably posited in our best scientific theories (Quine 1948, Putnam 1971).

So I have finished my second task: helping proponents of CEM see that they have a new argument to explore in their favor—an indispensability argument from the 2021 Nobel Prize in Economics. To further the dialectic, it is time for me to switch sides and help Lewisians undermine that argument.

4 Doing without Conditional Excluded Middle

I think that the above theory of causal inference can be reformulated and even generalized in a way that dispenses with CEM. This will be similar in spirit to what Field (2016) does to undermine the Quine-Putnam indispensability argument for mathematical realism when he reformulates Newtonian mechanics without real numbers.

4.1 First Step: The Rubin Causal Mode Made Stochastic

In the original game, everyone is only given a *single* card printed with '*if Take* = 1', whose face determines whether that person would, or would not, be cured under the treatment. But imagine that you are given not just one card printed with '*if Take* = 1' but a *deck* of such cards, in which 80% of the cards are printed with '*Cure* = 1' on their faces and the remaining 20% are printed with '*Cure* = 0'. Let this deck be well-shuffled, all faces down initially. Suppose that you took the treatment. Then Nature would *randomly* draw a card from this deck and flip it to reveal your medical result, and hence you would have an exactly 80% chance to be cured. So you could be cured and could be not cured—and thus it is not the case that you would be cured, nor is it the case that you would not cured. CEM is then rendered invalid, or so the Lewisians would argue. If randomly drawing a card from a deck does not sound chancy enough, replace it by measuring an observable in a quantum-mechanical system.

Let's generalize. In the original game, every individual is given four cards that answer four what-if questions, respectively:

(Q_1) What if one took the treatment?
(Q_2) What if one didn't take the treatment?
(Q_3) What if one were assigned to the treatment group?
(Q_4) What if one were assigned to the control group?

Now, let everyone's four cards be replaced by four decks, which answer the four what-if questions in this form: 'If individual i were ..., then i would have a probabilistic chance p to be ....' This p is a *counterfactual probability*, a probability under a counterfactual condition.

So we have a stochastic version of the Rubin causal model, thanks to the expansion pack for the card game. Deterministic outcomes are replaced by counterfactual probabilities, which can be used to redefine several concepts in the original Rubin causal model.

Each individual i still has an individual treatment effect (ITE), but now redefined as the difference between two counterfactual probabilities, or two proportions in decks of cards:

$$\begin{aligned} \mathrm{ITE}_i \;=_{\mathrm{df}}\; & \text{(i) the proportion of '}\mathit{Cure} = 1\text{' cards in } i\text{'s deck for '}\mathit{if\ Take} = 1\text{'} \\ & - \text{(ii) the proportion of '}\mathit{Cure} = 1\text{' cards in } i\text{'s deck for '}\mathit{if\ Take} = 0\text{'}. \end{aligned}$$

In the limiting case where each deck contains only one card, the ITE just defined reduces to the ITE defined earlier.

Every individual i now has a *degree of compliance* DC_i, defined by how one's chance of taking the treatment would change if one switched from the control group to the treatment group:

$$\begin{aligned} \mathrm{DC}_i \;=_{\mathrm{df}}\; & (a) \text{ the proportion of '}\mathit{Take} = 1\text{' cards in } i\text{'s deck for '}\mathit{if\ Assign} = 1\text{'} \\ & - (b) \text{ the proportion of '}\mathit{Take} = 1\text{' cards in } i\text{'s deck for '}\mathit{if\ Assign} = 0\text{'}. \end{aligned}$$

A degree of compliance can be positive, zero, or negative, corresponding to three subpopulations:

- If $\mathrm{DC}_i > 0$, then one is called a *complier* (in the general sense).
- If $\mathrm{DC}_i < 0$, then one is called a *defier* (in the general sense).
- If $\mathrm{DC}_i = 0$, then one is called an *indifferent-taker*.

The LATE is replaced by a more general concept: a weighted average of the individual treatment effects, in which everyone's weight w_i is proportional to that person's degree of compliance DC_i. So it is called the *degree-of-compliance-weighted average treatment effect* in the entire population, or DATE for short:

$$\begin{aligned} \mathrm{DATE} &=_{\mathrm{df}} \sum_i w_i \,\mathrm{ITE}_i \\ w_i &=_{\mathrm{df}} \frac{\mathrm{DC}_i}{\sum_j \mathrm{DC}_j} \end{aligned}$$

The denominator $\sum_j \mathrm{DC}_j$ is a normalizing factor, introduced only to ensure that the weights sum to 1.

Defiers, if any, have *negative* weights, which make it hard to interpret the weighted average. But let's follow the classic result in assuming that there are no defiers, so the

DATE receives no contributions from defiers. The DATE also receives no contributions from indifferent-takers, who carry zero weights by definition. It follows that only compliers make contributions to the DATE. The more compliant one is, the more weight one carries.

The present setting subsumes the original card game as a limiting case, in which all decks contain only one card. In this case, the compliers are all equally compliant, with a maximal degree of compliance: 100% minus 0%. The compliers then have equal weights in the DATE, which is turned into a simple average over the subpopulation of compliers, and thus degenerates to the LATE in this special case.

4.2 Second Step: Incorporating Stochastic Rubin into Causal Bayes

The application of instrumental variable estimation is often presented to rely on an assumption that can be formulated in plain English as follows: the assignment mechanism (to the treatment/control group) causally influences the medical outcome only through whether an individual takes the treatment. While this assumption has a standard probabilistic statement in the Rubin causal model (see the Assumption of *Instrumentality* in Appendix A), it can be restated in a way closer to the the plain English formulation, using the causal structure depicted in figure 1.

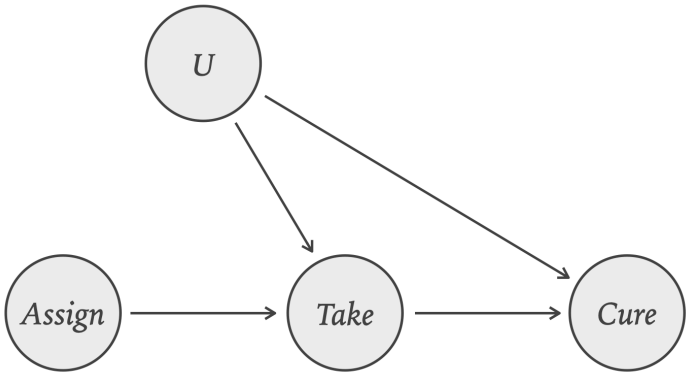

Figure 1: The causal structure in the instrumentality* assumption

In this causal graph, the *Assign* variable causally influences the *Cure* variable only through the *Take* variable. The confounding variable, written U, is the most fined-grained possible, whose value is the individual randomly drawn from the population.

To turn this causal graph into a causal Bayes net, it suffices to specify the probability distribution of each exogenous variable (such as U and $Assign$), and the probability distribution of each effect variable given its direct cause variables, as presented in figure 2. Everyone in the population has an equal chance of being chosen, so $\Pr(U = i) = 1/N$, where N is the population size. Once a person i is chosen, a coin is flipped to decide whether to assign that person to the treatment group or control group, so $\Pr(Assign = 1) = 1/2$, or more generally, $\Pr(Assign = 1)$ is a constant, independent of the individual chosen.

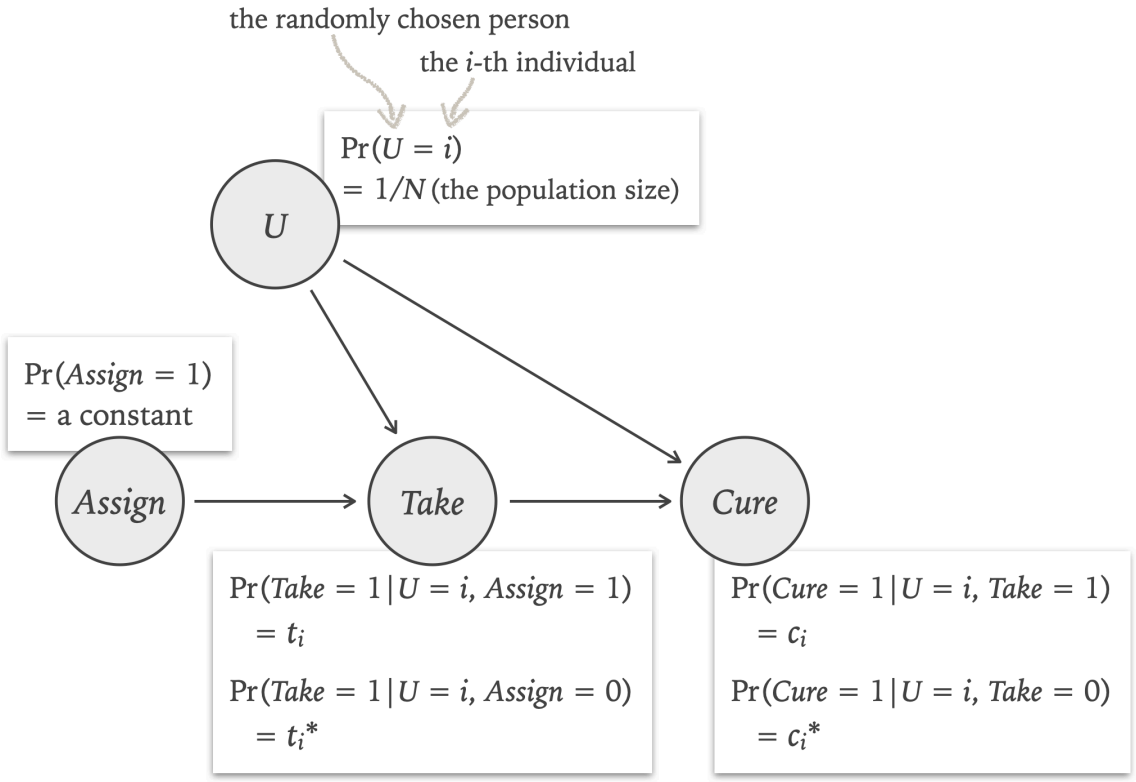

Figure 2: The causal Bayes net assumed in Theorem 2

Here is the crucial step: the conditional probabilities of effects given direct causes are identified with the appropriate counterfactual probabilities, which come from the stochastic version of Rubin causal model. In other words, the conditional probabilities c_i, c_i^*, t_i, and t_i^* as shown in the figure 2 are given by the counterfactual probabilities in the stochastic Rubin causal model as follows, using the language of the card game:

- c_i is the proportion of the '$Cure = 1$' cards in i's deck #1,
  i.e., the deck for '*if* $Take = 1$'.

- c_i^* is the proportion of the '*Cure* = 1' cards in i's deck #2.
- t_i is the proportion of the '*Take* = 1' cards in i's deck #3.
- t_i^* is the proportion of the '*Take* = 1' cards in i's deck #4.

This new causal modeling can be understood to work as follows. While the original Rubin causal model only allows an individual to have deterministic outcomes, my expansion pack for the card game updates it to allow for stochastic outcomes, with counterfactual probabilities that are plugged into the parameters of an appropriate causal Bayes net.

4.3 Main Result

Then we have a new identification result:

Theorem 2. *Suppose that the following assumptions hold:*

1. (Randomization) *Individuals in the population are randomly chosen with equal chances, and then the selected ones are randomly assigned to the treatment or control group with a constant bias* (*by, say, flipping a fair coin*).
2. (Instrumentality*) *The true causal model is a causal Bayes net with the causal structure in figure 1.*
3. (Existence of Compliers*) *There are compliers in the population, in the sense that someone's degree of compliance is positive.*
4. (No Defiers*) *There are no defiers in the population, in the sense that everyone's degree of compliance is nonnegative.*

Then the DATE *can be expressed solely by probabilities over the observable variables—Assign, Take, and Cure—without counterfactuals. Specifically, the* DATE *can be expressed as follows:*

$$\text{DATE} \;=\; \frac{\Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=0)}{\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=0)}\,.$$

See Appendix C for a proof. Assumptions 2-4 are labeled with asterisks in order to distinguish them from their counterparts in the original Rubin causal model as

stated in Appendix A: Assumption 2 (Instrumentality), Assumption 3 (Existence of Compliers), and Assumption 5 (No Defiers).

Given this more general approach to causal modeling, Lewisians can easily relax the assumption of CEM by replacing single cards with decks. Moreover, in this new theorem, the DATE equation's right-hand side remains identical to the LATE's in the classic result (Theorem 1). So, we can still use the same method of instrumental variable estimation designed originally for the LATE. But now we do this without assuming CEM—we are simply estimating the more general quantity DATE. This feature is important for undermining the indispensability argument. Medical and social scientists using the usual estimation method for the LATE under the logical principle of CEM can now be *reinterpreted* as estimating the DATE with the very same method, but without the assumption of CEM. This suggests that the new theorem allows us to preserve the Rubin causal model's successes in (apparently) estimating the LATE, while discarding the logical principle of CEM. We only need to reinterpret what is estimated in those successful applications.

This concludes my final task: helping Lewisians reject CEM by undermining the indispensability argument.

5 Closing

The Rubin causal model, with its underlying deductive logic and its ability to facilitate causal inference, warrants closer examination by philosophers. To this end, the previous discussion offered a card-game tutorial to introduce the model, and then developed a dialectic to connect it to some familiar philosophical ideas. The focus was on the ongoing debate surrounding a logical principle: Conditional Excluded Middle (CEM). I delved into both sides of the debate, in turn. First, I explored how the Rubin causal model could be used to construct a new argument for CEM—a Quine-Putnam indispensability argument. I then shifted gears and challenged this new argument, using causal Bayes nets to give a new theorem that seems to render CEM dispensable.

While my heart goes to the Lewisian side of the debate on CEM, that is only secondary for now. The real takeaway is how the dialectic highlights the intriguing potential of the Rubin causal model for philosophers. In fact, I see opportunities for both sides of the debate.

For proponents of CEM, the next step could be arguing that the very use of causal

Bayes nets presupposes CEM after all. On the other hand, opponents of CEM can delve deeper into the potential of causal Bayes nets as an improvement for the Rubin causal model, going beyond the instrumental variable estimation as discussed above. If health and social scientists can be persuaded to abandon CEM, it would represent an example in which scientific inquiry drives revisions in deductive logic—precisely the kind of example Quine (1951) envisaged. This would demonstrate the possibility of logic revision close to the realm of everyday concerns, like medicine or social issues—a far more relatable scenario than Putnam's (1968) case of quantum logic.

So much about deductive logic, but there is something for theorists of induction, too. When scientists justify inductive methods, they rely heavily on their context of inquiry, including background assumptions. Past discussions mostly focus on the background assumptions that are physical (Longino 1979, Christensen 1997), methodological, or ethical (Reiss 2020), rather than deductively logical. But is the logical principle CEM, for instance, needed to justify instrumental variable estimation? The deductive background deserves attention from theorists of induction.

There is even something for those more interested in modeling rather than inference. Consider the interplay between three approaches to causal modeling:

(1) nonparametric structural equation models (Pearl 2009),

(2) Rubin causal models (Rubin 1974),

(3) causal Bayes nets (Spirtes et al. 2000).

Pearl (2009) famously argues that the first two are basically equivalent and can produce everything that we get from the third approach—causal Bayes nets. But the new theorem suggests a different picture: in at least one application, namely the application to the LATE, causal Bayes nets seem to be able to generalize the Rubin causal models with an extended result (Theorem 2). So the questions remain: Which approach is more general? Which are equivalent, and in what sense? This would be a nice case study on intertheoretic relations. While initial steps have been taken by Markus (2021) and Weinberger (2023), they have not considered causal Bayes nets. Much more needs to be done.

The Rubin causal model clearly presents a rich landscape for further exploration.

References

Angrist, J. D. (1990) "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records", *American Economic Review*, 80, 313-336.

Christensen, D. (1997) "What Is Relative Confirmation?", *Noûs*, 31(3), 370-384.

Field, H. (2016) *Science without Numbers*, Oxford University Press.

Hájek, A. (MS) "Most Counterfactuals Are False", URL = <https://philarchive.org/rec/HJEMCA>

Hernán, M. A. & Robins, J. M. (2020) *Causal Inference: What If*, Chapman & Hall/CRC.

Imbens, G. W., & Angrist, J. (1994) "Identification and Estimation of Local Average Treatment Effects", *Econometrica* 62, 467-476.

Lewis, D. K. (1973) *Counterfactuals*, Blackwell.

Longino, H. E. (1979) "Evidence and Hypothesis: An Analysis of Evidential Relations", *Philosophy of Science*, 46(1), 35-56.

Mandelkern, M. (2022) "Modals and Conditionals", in Altshuler, D. (ed.) *Linguistics Meets Philosophy*, Oxford University Press, pp. 502-533.

Markus, K. A. (2021) "Causal Effects and Counterfactual Conditionals: Contrasting Rubin, Lewis and Pearl", *Economics & Philosophy*, 37(3), 441-461.

Pearl, J. (2009), *Causality*, Cambridge University Press.

Putnam, H. (1968) "Is Logic Empirical?", in Cohen, R. S. & Wartofsky, M. W. (eds.) *Boston Studies in the Philosophy of Science*, Vol. 5, D. Reidel: 216-241.

——— (1971) *Philosophy of Logic*, Routledge.

Quine, W. V. (1948) "On What There Is", *Review of Metaphysics*, 2(5): 21-38.

——— (1951) "Two Dogmas of Empiricism", *Philosophical Review*, 60: 20-43.

Reiss, J. (2020) "What Are the Drivers of Induction? Towards a Material Theory^{+}", *Studies in History and Philosophy of Science Part A*, 83, 8-16.

Rubin, D. B. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology* 66: 688-701.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000) *Causation, Prediction, and Search*, MIT Press.

Stalnaker, R. C. (1968) "A Theory of Conditionals", in Harper, W. L., Pearce, G. A., & Stalnaker, R. C. (eds.) *Ifs: Conditionals, Belief, Decision, Chance and Time*, Springer Netherlands: 41-55.

Weinberger, N. (2023) "Comparing Rubin and Pearl's Causal Modelling Frameworks: A Commentary on Markus (2021)", *Economics & Philosophy*, 39(3), 485-493.

Williams, J. R. G. (2010) *Defending Conditional Excluded Middle*, Noûs, 44(4), 650-668.

A Rigorous Presentation of the Rubin Causal Model

The second approach is basically the textbook-standard presentation of the Rubin causal model. The formalism builds on elementary probability theory: ordinary variables are upgraded to potential outcomes, i.e. variables under counterfactual, subjunctive conditions.

A.1 Counterfactuals & Potential Outcomes

Let $Take_i = 1$ expresses the proposition that the individual i takes the treatment. Similarly for $Cure_i = 1$ and $Assign_i = 1$. To that notation we can add superscripts to express counterfactuals, such as the following:

- $Cure_i^{Take_i=1} = 1$ means that the individual i would be cured if i took the treatment.
- $Cure_i^{Take_i=1, Take_j=0} = 0$ means that the individual i would not be cured if i took the treatment but another individual, j, did not.

The variables like $Cure_i^{Take_i=1}$ and $Cure_i^{Take_i=1, Take_j=0}$ are just ordinary variables under a counterfactual, subjunctive condition. More generally, X_i^{C}, called a *potential outcome*, denotes the value of variable X that individual i would have under the subjunctive condition C. I love the *superscript* notation, for it makes vivid the idea of "*under* a condition", although there are other notations. For example, $Cure_i^{Take_i=1}$ is often written as $Cure_i(1)$ when the context makes clear what's intended.

A.2 Assumptions about Potential Outcomes

There are also substantive, empirical assumptions:

> **Assumption 1 (Stable Unit Treatment Value, or SUTVA).** The values of the variables of each individual (or unit) are determined in a way independent of the values of the variables of any other individuals. That is, for any variable X_i and any conditions $\mathrm{C}_1, \ldots, \mathrm{C}_n$ about individuals from 1 to n, we have $X_i^{\mathrm{C}_1,\ldots,\mathrm{C}_n} = X_i^{\mathrm{C}_i}$, which omits any references to individuals other than i in the subjunctive condition.

So, to think about whether i would be cured under various conditions, it suffices to consider potential outcomes of the form $Cure_i^{Assign_i=a, Take_i=t}$, which makes no reference to any other individuals in the subjunctive condition. This allows us to consider only a small number of potential outcomes. The next assumption allows us to drop more terms in the subjunctive conditions and thus consider an even smaller number of potential outcomes:

> **Assumption 2 (Instrumentality).** For each individual i, $Assign_i$ is an instrumental variable in the following sense: the value of $Cure_i$ is determined once the value of $Take_i$ is determined, independently of the value of $Assign_i$. That is, $Cure_i^{Assign_i=a, Take_i=t} = Cure_i^{Take_i=t}$, which omits the assignment $Assign_i = a$ in the subjunctive condition.

Thanks to the above two assumptions, now we only need to consider just four potential outcomes for each individual i: $Cure_i^{Take_i=1}$, $Cure_i^{Take_i=0}$, $Take_i^{Assign_i=1}$, $Take_i^{Assign_i=0}$. Those four variables correspond to the four cards that i has in the game. In fact, the faces of the four cards are printed with the values taken by those four potential outcomes, respectively. When we decide to model each individual by a set of such four cards, we are already committed to the assumptions of SUTVA and Instrumentality.

Since the four cards have been replaced formally by four potential outcomes, the concepts presented above can be redefined formally accordingly. Every individual i is classified into one of the following categories:

$$\begin{aligned}
\mathit{Complier}_i &\Leftrightarrow_{\text{df}} \mathit{Take}_i^{\mathit{Assign}_i=0} = 0 \wedge \mathit{Take}_i^{\mathit{Assign}_i=1} = 1\,; \\
\mathit{Defier}_i &\Leftrightarrow_{\text{df}} \mathit{Take}_i^{\mathit{Assign}_i=0} = 1 \wedge \mathit{Take}_i^{\mathit{Assign}_i=1} = 0\,; \\
\mathit{Always\text{-}Taker}_i &\Leftrightarrow_{\text{df}} \mathit{Take}_i^{\mathit{Assign}_i=0} = 1 \wedge \mathit{Take}_i^{\mathit{Assign}_i=1} = 1\,; \\
\mathit{Never\text{-}Taker}_i &\Leftrightarrow_{\text{df}} \mathit{Take}_i^{\mathit{Assign}_i=0} = 0 \wedge \mathit{Take}_i^{\mathit{Assign}_i=1} = 0\,.
\end{aligned}$$

The ITE (individual treatment effect) on an individual i is defined by

$$\text{ITE}_i \;=_{\text{df}}\; \mathit{Cure}_i^{\mathit{Take}_i=1} - \mathit{Cure}_i^{\mathit{Take}_i=0}\,.$$

The local average treatment effect (on the compliers) is defined by

$$\text{LATE} \;=_{\text{df}}\; \frac{1}{\#\{i : \mathit{Complier}_i\}} \sum_{i:\, \mathit{Complier}_i} \text{ITE}_i\,.$$

To make this well-defined, the denominator has to be assumed to be nonzero:

Assumption 3 (Existence of Compliers)
$\mathit{Complier}_i$ for some individual i.

A.3 Probabilistic Assumptions

Some empirical assumptions are less general, meant to serve the purposes of instrumental variable estimation.

Let the subscript-free notation $\Pr\big(\mathit{Cure}^{\mathit{Take}=0} = 1\big)$ denote the probability of randomly choosing an individual who would be cured (even) without taking the treatment. If everyone has an equal probability $1/N$ to be chosen, where N is the population size, then $\Pr\big(\mathit{Cure}^{\mathit{Take}=0} = 1\big)$ is identical to the proportion of those who would be cured without taking the treatment—that is, the individuals i with the feature $\mathit{Cure}_i^{\mathit{Take}_i=0} = 1$. Similarly, given that a randomly chosen individual is assigned to the treatment group, the probability of that person taking the treatment is denoted by $\Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 1)$.

Assumption 4 (Random Choice). Everyone in the population has an

equal probability of being chosen for participation of the experiment. So the actual frequency distribution of the four potential outcomes in the population is the same as the probability distribution of those variables. In other words:

the proportion of those in the population with features
$Take^{Assign=0} = x,\ Take^{Assign=1} = y,\ Cure^{Take=0} = z$, and $Cure^{Take=1} = u$

$= \Pr\big(Take^{Assign=0} = x,\ Take^{Assign=1} = y,\ Cure^{Take=0} = z,\ Cure^{Take=1} = u\big)$

Lemma 1. *By the above assumptions, the* LATE *can be expressed probabilistically:*

$$\text{LATE} \ = \ \Pr\big(Cure^{Take=1} = 1 \,|\, Complier\big) \ - \ \Pr\big(Cure^{Take=0} = 1 \,|\, Complier\big) \ .$$

Proof. Calculate the LATE as follows; Existence of Compliers is assumed throughout to make all denominators nonzero:

$$\begin{aligned}
&\text{LATE} \\
&= \ \frac{1}{\#\{i : Complier_i\}} \sum_{i:\, Complier_i} \text{ITE}_i \\
&= \ \frac{1}{\#\{i : Complier_i\}} \sum_{i:\, Complier_i} \big(Cure_i^{Take_i=1} - Cure_i^{Take_i=0}\big) \\
&= \ \frac{\sum_{i:\, Complier_i} Cure_i^{Take_i=1}}{\#\{i : Complier_i\}} \ - \ \frac{\sum_{i:\, Complier_i} Cure_i^{Take_i=0}}{\#\{i : Complier_i\}} \\
&= \ \frac{\#\{i : Cure_i^{Take_i=1} = 1 \wedge Complier_i\}}{\#\{i : Complier_i\}} \ - \ \frac{\#\{i : Cure_i^{Take_i=1} = 0 \wedge Complier_i\}}{\#\{i : Complier_i\}} \\
&= \ \Pr\big(Cure^{Take=1} = 1 \,|\, Complier\big) \ - \ \Pr\big(Cure^{Take=0} = 1 \,|\, Complier\big) \ .
\end{aligned}$$

The last step applies the assumption of Random Choice. □

Standard textbooks often define the LATE by the formula in the preceding lemma or by a slight variant that replaces probabilities by expected values. But I believe that the present approach enhances conceptual clarity: the LATE is defined solely in terms of just the individuals and their properties in the population, independently of the probabilities that come (partly or entirely) from the randomization mechanism that

scientists set up in order to study the population.

The four assumptions made above, (i) SUTVA, (ii) Instrumentality, (iii) Existence of Compliers, and (iv) Random Choice have been used to define some concepts and prove the above probabilistic formula for the LATE. In fact, those assumptions have finished their jobs and will no longer be cited in this section. There are additional assumptions, which are needed to prove more lemmas.

A.4 Auxiliary Assumptions

An assumption has been made regarding a subpopulation (Existence of Compliers); here is an assumption about another subpopulation:

> **Assumption 5 (No Defiers)** $Defier_i$ for no individual i.

In addition to the assumption of Random Choice, there is another probabilistic assumption:

> **Assumption 6 (Random Assignment).** Any individual, once chosen for participation of the experiment, has a fixed probability (say 50%) of being assigned to the treatment/control group, independently of their identity. So, *Assign* is probabilistically independent of the set of all the four potential outcomes in use, $Take^{Assign=0}$, $Take^{Assign=1}$, $Cure^{Take=0}$, and $Cure^{Take=1}$; or in symbols:

$$\begin{aligned}
&\Pr\big(Take^{Assign=0} = x,\ Take^{Assign=1} = y,\ Cure^{Take=0} = z,\ Cure^{Take=1} = u\big) \\
&= \Pr\big(Take^{Assign=0} = x,\ Take^{Assign=1} = y,\ Cure^{Take=0} = z,\ Cure^{Take=1} = u \,|\, Assign = 0\big) \\
&= \Pr\big(Take^{Assign=0} = x,\ Take^{Assign=1} = y,\ Cure^{Take=0} = z,\ Cure^{Take=1} = u \,|\, Assign = 1\big)\ .
\end{aligned}$$

There are two assumptions about the logic of counterfactuals:

> **Assumption 7 (Consistency/Centering).** It must be that
>
> $$X_i = x \ \Rightarrow\ \big(Y_i^{X_i=x} = y \ \Leftrightarrow\ Y_i = y\big)\,;$$
>
> that is, an antecedent $X_i = x$ in a counterfactual is redundant if it turns out to be true.

Assumption 8 (Conditional Excluded Middle). Suppose that Y_i is a binary variable. Then the counterfactual variable $Y_i^{X_i=x}$, understood to denote the value that Y_i would have under the subjunctive condition that $X_i = x$, is still a binary variable; that is,

$$Y_i^{X_i=x} = 0 \ \vee \ Y_i^{X_i=x} = 1\,.$$

For example, it is assumed that $\mathit{Cure}_i^{\mathit{Take}_i=1} = 0 \ \vee \ \mathit{Cure}_i^{\mathit{Take}_i=1} = 1$, which says that i would be cured under the treatment, or i would not be cured under the treatment (with the same if-clause).

Then we have this classic result:

Theorem 1 (Imbens & Angrist 1994)**.** *Under the eight assumptions stated in this section, we have:*

$$\text{LATE} \ = \ \frac{\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 1) \ - \ \Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 0)}{\Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 1) \ - \ \Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 0)}\,.$$

B Proof of Theorem 1

I will begin by explaining the idea of proof. The lemmas in used will then be presented and proved.

B.1 Idea of Proof: What Can We Learn from Observations without Manipulations?

By pure observations, we can register the proportion of those taking the treatment in the treatment (or control) group, and thereby estimate the following conditional probabilities about observables:

$$\begin{aligned} \Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 0) \ &= \ ? \\ \Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 1) \ &= \ ? \end{aligned}$$

Similarly, we can register the proportion of the cured among those who takes (or does not take) the treatment in the treatment (or control) group, and thereby estimate the

following conditional probabilities about observables:

$$\begin{aligned}
\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 0, \mathit{Take} = 0) &= ? \\
\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 0, \mathit{Take} = 1) &= ? \\
\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 1, \mathit{Take} = 0) &= ? \\
\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 1, \mathit{Take} = 1) &= ?
\end{aligned}$$

If any information about counterfactuals is to be learned from pure observations of the three observables—*Assign*, *Take*, and *Cure*—this information must be in principle extractable from the above six conditional probabilities. For those conditional probabilities suffice to uniquely determine a joint probability distribution over the three observables:

$$\begin{aligned}
&\Pr(\mathit{Assign} = a, \mathit{Take} = t, \mathit{Cure} = c) \\
&= \Pr(\mathit{Cure} = c \,|\, \mathit{Take} = t, \mathit{Assign} = a) \cdot \Pr(\mathit{Take} = t \,|\, \mathit{Assign} = a) \cdot \underbrace{\Pr(\mathit{Assign} = a)}_{= 0.5} .
\end{aligned}$$

The question is what information about counterfactuals can be extracted from just those six probabilities.

It turns out that those six conditional probabilities about observables, once known or estimated, can tell us a lot about counterfactuals. Assuming no defiers, the first two can tell us the proportions of some subpopulations (which will be stated officially in lemma 2 below):

$$\Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 0) = \Pr(\mathit{Always\text{-}Taker}) \tag{1}$$

$$\Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 1) = \Pr(\mathit{Complier} \vee \mathit{Always\text{-}Taker}) \tag{2}$$

Assuming no defiers (again), it follows that the proportions of three subpopulations

can all be estimated observationally (to be stated officially in lemma 3 below):

$$\begin{aligned}
\Pr(\textit{Always-Taker}) &= \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 0)\ ; && (3)\\
\Pr(\textit{Never-Taker}) &= 1 - \Pr(\textit{Complier} \vee \textit{Always-Taker}) \\
&= 1 - \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 1) \\
&= \Pr(\textit{Take} = 0 \,|\, \textit{Assign} = 1)\ ; && (4)\\
\Pr(\textit{Complier}) &= \Pr(\textit{Complier} \vee \textit{Always-Taker}) - \Pr(\textit{Always-Taker}) \\
&= \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 1) - \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 0)\ . && (5)
\end{aligned}$$

Every term on the right side can be estimated by observations.

The remaining four conditional probabilities also tell us something substantive about counterfactuals (which will be stated officially in lemma 4 below):

$$\begin{aligned}
\Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 0, \textit{Take} = 0) &= \Pr\big(\textit{Cure}^{\textit{Take}=0} = 1 \,|\, \textit{Complier} \vee \textit{Never-Taker}\big) && (6)\\
\Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 0, \textit{Take} = 1) &= \Pr\big(\textit{Cure}^{\textit{Take}=1} = 1 \,|\, \textit{Always-Taker}\big) && (7)\\
\Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 1, \textit{Take} = 0) &= \Pr\big(\textit{Cure}^{\textit{Take}=0} = 1 \,|\, \textit{Never-Taker}\big) && (8)\\
\Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 1, \textit{Take} = 1) &= \Pr\big(\textit{Cure}^{\textit{Take}=1} = 1 \,|\, \textit{Complier} \vee \textit{Always-Taker}\big) && (9)
\end{aligned}$$

This much is already enough to let us observationally estimate the two terms in the probabilistic expression for the LATE (as in lemma 1):

$$\text{LATE} \;=\; \Pr\big(\textit{Cure}^{\textit{Take}=1} = 1 \,|\, \textit{Complier}\big) \;-\; \Pr\big(\textit{Cure}^{\textit{Take}=0} = 1 \,|\, \textit{Complier}\big)\ ,$$

To estimate the first term $\Pr\big(\textit{Cure}^{\textit{Take}=1} = 1 \,|\, \textit{Complier}\big)$, consider (7) and (9): they tell us the proportion of those with $\textit{Cure}^{\textit{Take}=1} = 1$ (who would be cured under the treatment) in some subpopulations: in the smaller subpopulation comprising just always-takers (7), and in the more inclusive subpopulation comprising the compliers and the always-takers (9). We can then derive the proportion of those with $\textit{Cure}^{\textit{Take}=1} = 1$ in the *difference* between those two subpopulations, i.e., in the subpopulation comprising just the compliers. This can be done by a routine procedure in elementary probability theory—by solving for the only unknown in the following

equation, where B and B' are mutually exclusive:

$$\Pr(A \,|\, B \vee B') \;=\; \underbrace{\Pr(A \,|\, B)}_{\text{unknown}} \tfrac{\Pr(B)}{\Pr(B)+\Pr(B')} \;+\; \Pr(A \,|\, B') \tfrac{\Pr(B')}{\Pr(B)+\Pr(B')}$$

To be more specific, let A be $\mathit{Cure}^{\mathit{Take}=1} = 1$, B be *Complier*, and B' be *Always-Taker*. Except for the term marked as unknown, which is $\Pr\big(\mathit{Cure}^{\mathit{Take}=1} = 1 \,|\, \mathit{Complier}\big)$, every other term in the above formula can be expressed by the six conditional probabilities about observables and, thus, estimated by observational data. This explains how a bit of algebra allows the unknown term $\Pr\big(\mathit{Cure}^{\mathit{Take}=1} = 1 \,|\, \mathit{Complier}\big)$ to be expressed by the probabilities about observables.

Once we know how to express the first term $\Pr\big(\mathit{Cure}^{\mathit{Take}=1} = 1 \,|\, \mathit{Complier}\big)$ by the probabilities about observables, we can apply the same trick to express the second term $\Pr\big(\mathit{Cure}^{\mathit{Take}=0} = 1 \,|\, \mathit{Complier}\big)$ also by the probabilities about observables. Then, by taking their difference, we get a formula that expresses the LATE by the probabilities of observables, too. This formula might look a bit ugly initially. But, at this point, it only takes a bit of calculations in elementary probability theory to obtain a beautiful formula as stated in Theorem 1:

$$\text{LATE} \;=\; \frac{\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 1) \;-\; \Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 0)}{\Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 1) \;-\; \Pr(\mathit{Take} = 1 \,|\, \mathit{Assign} = 0)}\,.$$

This finishes the idea of proof.

B.2 Lemmas and Proofs

Although the four groups of people (compliers, defiers, always-takers, and never-takers) are defined in terms of counterfactuals, we can still identify them in the right conditions. Think about an individual i in the control group: If i takes the treatment, then i can only be an always-taker or a defier (rather than a never-taker or complier), and hence i must be an always-taker (for defiers have been assumed to be absent). Conversely, if i is an always-taker, i must take the treatment (despite being assigned to the control group). It follows that, within the control group, those taking the treatment are *exactly* those being always-takers. This result is important, for it implies that we can identify and observe a sample of always-takers by looking at the people taking the treatment in the control group. This result is presented as clause (i) of Lemma 2; similar results are in the other clauses:

Lemma 2. *The reference to individuals i is surpassed throughout for the sake of readability. Under the above assumptions, being assigned to the control group, $Assign = 0$, implies*

(i) $Take = 1 \;\Leftrightarrow\; Always\text{-}Taker$,
(ii) $Take = 0 \;\Leftrightarrow\; Complier \vee Never\text{-}Taker$.

Similarly, being assigned to the treatment group, $Assign = 1$, implies

(iii) $Take = 0 \;\Leftrightarrow\; Never\text{-}Taker$,
(iv) $Take = 1 \;\Leftrightarrow\; Complier \vee Always\text{-}Taker$.

It follows immediately that

$$\Pr(Take = 1 \,|\, Assign = 0) \;=\; \Pr(Always\text{-}Taker)\;; \tag{10}$$
$$\Pr(Take = 1 \,|\, Assign = 1) \;=\; \Pr(Complier \vee Always\text{-}Taker)\;. \tag{11}$$

Proof. Prove part (i) as follows:

$$\begin{aligned}
& Assign = 0 \wedge Always\text{-}Taker \\
\Leftrightarrow\; & Assign = 0 \wedge (Always\text{-}Taker \vee Defier) && \text{by No Defiers} \\
\Leftrightarrow\; & Assign = 0 \wedge \\
& \Big[\big(Take^{Assign=0} = 1 \wedge Take^{Assign=1} = 1\big) \vee \\
& \quad \big(Take^{Assign=0} = 1 \wedge Take^{Assign=1} = 0\big) \Big] && \text{by definitions} \\
\Leftrightarrow\; & Assign = 0 \wedge \\
& \Big[Take^{Assign=0} = 1 \wedge \\
& \quad \underbrace{\big(Take^{Assign=1} = 1 \;\vee\; Take^{Assign=1} = 0\big)}_{\text{redundant by Conditional Excluded Middle}} \Big] && \text{by De Morgan Rule} \\
\Leftrightarrow\; & Assign = 0 \wedge Take^{Assign=0} = 1 && \text{by Conditional Excluded Middle} \\
\Leftrightarrow\; & Assign = 0 \wedge Take = 1 && \text{by Consistency}
\end{aligned}$$

Part (ii) can be obtained from part (i) by taking the contraposition of the equivalence '$\Leftrightarrow$' and by applying the assumption that there are no defiers. The remaining parts, (iii) and (iv), can be proved similarly. □

The preceding lemma helps to prove the next one:

Lemma 3. *The proportions of the three subpopulations can be expressed by the conditional probabilities of the three observables:*

$$\Pr(\textit{Always-Taker}) = \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 0)\ ; \tag{12}$$

$$\Pr(\textit{Never-Taker}) = \Pr(\textit{Take} = 0 \,|\, \textit{Assign} = 1)\ ; \tag{13}$$

$$\Pr(\textit{Complier}) = \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 1) - \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 0) \tag{14}$$

Proof. The probability of drawing a always-taker can be expressed as follows:

$$\begin{aligned}
&\Pr(\textit{Always-Taker}) \\
&= \Pr(\textit{Always-Taker} \,|\, \textit{Assign} = 0) && \text{by Random Assignment} \\
&= \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 0) && \text{by Lemma 2}
\end{aligned}$$

The probability of drawing a never-taker can be expressed similarly:

$$\begin{aligned}
&\Pr(\textit{Never-Taker}) \\
&= \Pr(\textit{Never-Taker} \,|\, \textit{Assign} = 1) && \text{by Random Assignment} \\
&= \Pr(\textit{Take} = 0 \,|\, \textit{Assign} = 1) && \text{by Lemma 2}
\end{aligned}$$

Since there are no defiers, the probability of drawing a complier is equal to 1 minus the two probabilities in the above; that is:

$$\begin{aligned}
&\Pr(\textit{Complier}) \\
&= 1 - \Pr(\textit{Always-Taker}) - \Pr(\textit{Never-Taker}) && \text{by No Defiers} \\
&= 1 - \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 0) - \Pr(\textit{Take} = 0 \,|\, \textit{Assign} = 1)
\end{aligned}$$

So we are done. □

Lemma 4.

$$\begin{aligned}
\Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 0, \textit{Take} = 0) &= \Pr\big(\textit{Cure}^{\textit{Take}=0} = 1 \,|\, \textit{Complier} \vee \textit{Never-Taker}\big) \\
\Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 0, \textit{Take} = 1) &= \Pr\big(\textit{Cure}^{\textit{Take}=1} = 1 \,|\, \textit{Always-Taker}\big) \\
\Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 1, \textit{Take} = 0) &= \Pr\big(\textit{Cure}^{\textit{Take}=0} = 1 \,|\, \textit{Never-Taker}\big) \\
\Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 1, \textit{Take} = 1) &= \Pr\big(\textit{Cure}^{\textit{Take}=1} = 1 \,|\, \textit{Complier} \vee \textit{Always-Taker}\big)
\end{aligned}$$

Proof. The first equation can be proved as follows.

$$
\begin{aligned}
&\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 0, \mathit{Take} = 0) \\
&= \Pr\big(\mathit{Cure}^{\mathit{Take}=0} = 1 \,|\, \mathit{Assign} = 0, \mathit{Take} = 0\big) && \text{by Consistency} \\
&= \Pr\big(\mathit{Cure}^{\mathit{Take}=0} = 1 \,|\, \mathit{Assign} = 0, (\mathit{Complier} \vee \mathit{Never\text{-}Taker})\big) && \text{by Lemma 2} \\
&= \Pr\big(\mathit{Cure}^{\mathit{Take}=0} = 1 \,|\, \mathit{Complier} \vee \mathit{Never\text{-}Taker}\big) && \text{by Random Assignment}
\end{aligned}
$$

The second equation can be proved with the same strategy (in fact, the annotations have the same contents, even in the same order):

$$
\begin{aligned}
&\Pr(\mathit{Cure} = 1 \,|\, \mathit{Assign} = 0, \mathit{Take} = 1) \\
&= \Pr\big(\mathit{Cure}^{\mathit{Take}=1} = 1 \,|\, \mathit{Assign} = 0, \mathit{Take} = 1\big) && \text{by Consistency} \\
&= \Pr\big(\mathit{Cure}^{\mathit{Take}=1} = 1 \,|\, \mathit{Assign} = 0, \mathit{Always\text{-}Taker}\big) && \text{by Lemma 2} \\
&= \Pr\big(\mathit{Cure}^{\mathit{Take}=1} = 1 \,|\, \mathit{Always\text{-}Taker}\big) && \text{by Random Assignment}
\end{aligned}
$$

The third and fourth equations can be proved with the same strategy. □

The above lemmas cover all the conditional probabilities of observables discussed in the previous subsection. As explained toward the end of the preceding subsection (Appendix B.1), the rest is just a bit of calculations in elementary probability theory in order to finish a proof of theorem 1.

C Proof of Theorem 2

Proof. Thanks to the background provided in the precious Appendix, each individual i has an individual treatment effect

$$\mathrm{ITE}_i \;=\; c_i - c_i^* \,.$$

with a degree of compliance

$$\mathrm{DC}_i \;=\; t_i - t_i^* \,.$$

Hence the DATE is given by:

$$\begin{aligned}\text{DATE} &= \sum_i \underbrace{\left(\frac{\text{DC}_i}{\sum_j \text{DC}_j}\right)}_{=w_i} \text{ITE}_i \\ &= \sum_i \left(\frac{t_i - t_i^*}{\sum_j \left(t_j - t_j^*\right)}\right)\left(c_i - c_i^*\right),\end{aligned}$$

which is well-defined, i.e., the denominator is nonzero, thanks to the assumption of the Existence of Compliers*. The goal is to verify the following equation:

$$\text{DATE} \stackrel{?}{=} \frac{\Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 1) \,-\, \Pr(\textit{Cure} = 1 \,|\, \textit{Assign} = 0)}{\Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 1) \,-\, \Pr(\textit{Take} = 1 \,|\, \textit{Assign} = 0)}\,.$$

The terms on the right-hand side are to be calculated in turn. I will leverage a defining feature of the causal Bayes net, the *Causal Markov Assumption*: every variable is probabilistically independent of its non-descendants (non-effects) given its parents

(direct causes). Start with the first term in the numerator:

$$
\begin{aligned}
&\Pr(\mathit{Cure}=1 \mid \mathit{Assign}=1) \\
&= \sum_{i,j} \Big(\Pr(\mathit{Cure}=1 \mid \mathit{Take}=j, U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(\mathit{Take}=j \mid U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(U=i \mid \mathit{Assign}=1) \Big) && \text{by Chain Rule} \\
&= \sum_{i,j} \Big(\Pr(\mathit{Cure}=1 \mid \mathit{Take}=j, U=i, \cancel{\mathit{Assign}=1}) \\
&\qquad \times \Pr(\mathit{Take}=j \mid U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(U=i \mid \cancel{\mathit{Assign}=1}) \Big) && \text{by Causal Markov} \\
&= \sum_{i} \Big(\Pr(\mathit{Cure}=1 \mid \mathit{Take}=1, U=i) \\
&\qquad \times \Pr(\mathit{Take}=1 \mid U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(U=i) \Big) \\
&\quad + \sum_{i} \Big(\Pr(\mathit{Cure}=1 \mid \mathit{Take}=0, U=i) \\
&\qquad \times \Pr(\mathit{Take}=0 \mid U=i, \mathit{Assign}=1) \\
&\qquad \times \Pr(U=i) \Big) \\
&= \sum_{i} \Big(c_i \, t_i \, \tfrac{1}{N} \Big) \; + \; \sum_{i} \Big(c_i^* \, (1-t_i) \, \tfrac{1}{N} \Big) \\
&= \tfrac{1}{N} \sum_{i} \Big(c_i \, t_i + c_i^* \, (1-t_i) \Big) .
\end{aligned}
$$

Similarly for the second term in the numerator:

$$
\begin{aligned}
&\Pr(\mathit{Cure}=1 \mid \mathit{Assign}=0) \\
&= \tfrac{1}{N} \sum_{i} \Big(c_i \, t_i^* + c_i^* \, (1-t_i^*) \Big) .
\end{aligned}
$$

Now calculate the first term in the denominator:

$$\begin{aligned}
&\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=1) \\
&= \sum_i \Big(\Pr(\mathit{Take}=1 \,|\, U=i, \mathit{Assign}=1) \cdot \Pr(U=i \,|\, \underbrace{\cancel{\mathit{Assign}=1}}_{\text{by Causal Markov}}) \Big) \\
&= \sum_i t_i \, \tfrac{1}{N} \\
&= \tfrac{1}{N} \sum_i t_i \,.
\end{aligned}$$

Similarly for the second term in the denominator:

$$\begin{aligned}
&\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=0) \\
&= \tfrac{1}{N} \sum_i t_i^* \,.
\end{aligned}$$

To finish off, plug the four terms just calculated into the following:

$$\begin{aligned}
&\frac{\Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Cure}=1 \,|\, \mathit{Assign}=0)}{\Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=1) \;-\; \Pr(\mathit{Take}=1 \,|\, \mathit{Assign}=0)} \\
&= \frac{\cancel{\frac{1}{N}} \sum_i \Big(c_i \, t_i + c_i^* \, (1 - t_i) \Big) - \cancel{\frac{1}{N}} \sum_i \Big(c_i \, t_i^* + c_i^* \, (1 - t_i^*) \Big)}{\cancel{\frac{1}{N}} \sum_i t_i - \cancel{\frac{1}{N}} \sum_i t_i^*} \\
&= \frac{\sum_i \big(t_i - t_i^*\big)\big(c_i - c_i^*\big)}{\sum_j \big(t_j - t_j^*\big)} \\
&= \sum_i \left(\frac{t_i - t_i^*}{\sum_j \big(t_j - t_j^*\big)} \right) \big(c_i - c_i^*\big) \\
&= \mathrm{DATE} \,,
\end{aligned}$$

as desired. □

It is interesting to note that this proof makes no use of the assumption of No Defiers*, which only serves to make the weights nonnegative and, thus, interpretable as weights in a weighted average.