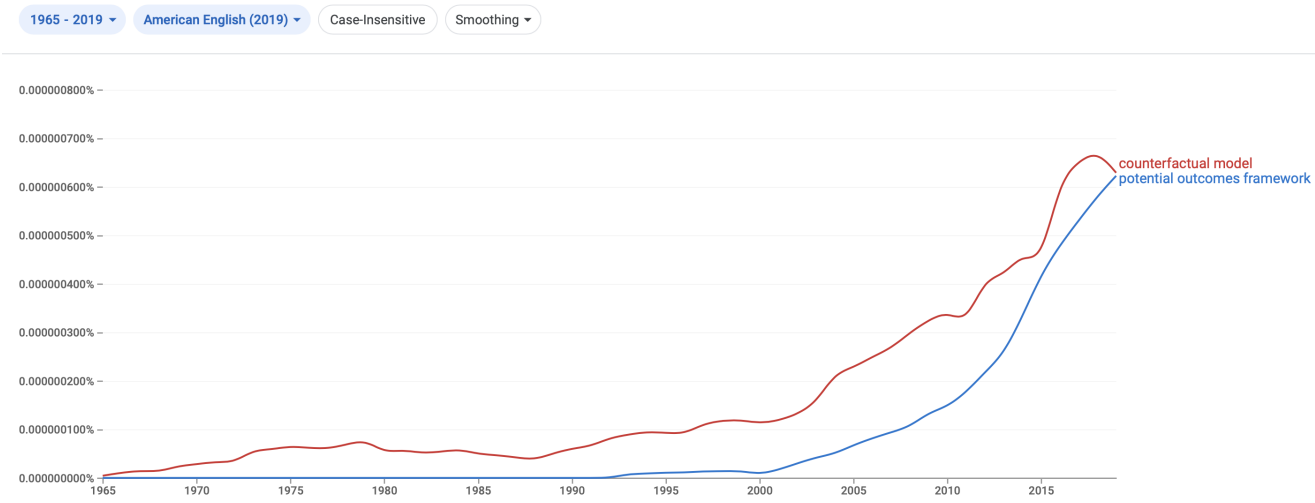
2 Potential Outcomes

Coding time

The sorting hat

Potential outcomes framework

The counterfactual model (aka the potential outcomes framework) has its origins in early work on experimental design by Neyman (1923), Fisher (1935) (both were concerned with agricultural experiments), Cochran and Cox (1950), Kempthorne (1952), and Cox (1958). Donald Rubin (1974) formalized the counterfactual model for causal analysis of observational data. In the statistics tradition, the model is often referred to as the potential outcomes framework, with reference to potential yields from Neyman's work in agricultural statistics. The counterfactual model also has roots in the economics literature (Roy 1951; Quandt 1972) (the so-called Roy Model), with important subsequent contributions by James Heckman, Charles Manski, and others. The model is now dominant in both statistics and economics, and it is being used with increasing frequency in sociology, psychology, and political science.



In a simple binary case, the *potential outcomes framework* (or the *counterfactual model*) assumes the existence of two well-defined *causal states* to which all members of the population of interest could be exposed. These two states are usually called *treatment* and *control*. In a non-binary case, one can refer to the alternative states as *alternative treatments*.

The notion of a *well-defined state* is not always trivial. Some examples have well-defined states, and others do not. Consider the job-training example. The two causal states for each individual are "received the job training" (treatment) and "did not receive the job training" (control). In the hospital example, the two causal states are "went to a hospital" (treatment) and "did not go to a hospital" (control). In the class size example, the two causal states are "assigned to a smaller class" (treatment) and "assigned to a regular class" (control). In the human capital example, the two causal states are "received a college degree" (treatment) and "did not receive a college degree" (control).

Note

We have to be careful in assuming that the meaning of what constitutes a treatment (job training, hospital, college) is the same for everyone. Suppose, e.g., there are good and bad colleges. Do we really want to mix them together? We will come back to that point later.

Other examples have less clearly defined causal states. Consider the relationship between socioeconomic status and political participation. There are many well-defined causal effects. E.g., the effect of having obtained at least a college degree on the frequency of voting in local elections. Or the effect of having a family income greater than some cutoff value on the amount of money donated to political campaigns. Well-defined causal states exist for these narrow causal effects, but it is not clear that well-defined causal states exist for the general concepts of "socioeconomic status" and "political participation."

When we are defining causal states, we are making a *ceteris paribus* assumption. In some cases, the assumption might be unrealistic if it rules out other changes that *must* occur at the same time. Notice a relationship to the ideal experiment from the previous lecture. If your ceteris paribus assumption cannot hold even in the ideal experiment, your causal relationship of interest is not identified.

Suppose our research question is whether starting school later (say, at age 7 vs. at age 6) improves a student's academic performance. Imagine we can randomize students into the treatment (start at 7) and control groups (start at 6). Let's look at their performance in the first grade. Students who start at 7 are likely to get better grades than students who start at 6 simply because of the maturation effect (7yo are older than 6yo). Ok, how about we look at the kids of the same age but in different grades. We will test the treatment group in the first grade and the control group in the second grade, when both groups are 7yo. But now the control group spent more time in school, so it might do better. There seems to be no way to identify the start-age effect separately from the maturation and time-in-school effects. The ceteris paribus does not hold.

But let's assume that our causal states are well-defined. Now we can define *potential outcome random variables* over all individuals in the population of interest. For a binary case, we can denote them as Y^1 (potential outcome in the treatment state) and Y^0 (potential outcome in the control state). We will denote y_i^1 as the potential outcome in the treatment state for individual i, and y_i^0 as the potential outcome in the control state for individual i. The individual-level causal effect of the treatment is then

$$\delta_i = y_i^1 - y_i^0.$$

> **Note**
>
> Individual-level causal effects can be defined in different ways, for example, as the ratio y_i^1/y_i^0. However, the above definition of the individual-level effect can easily accommodate ratios by redefining the outcomes as logarithms.

We define a causal exposure variable, D, which takes on two values: D is equal to 1 for members of the population who are exposed to the treatment state (the treatment group) and equal to 0 for members of the population who are exposed to the control state (the control group). An important question we have to ask ourselves is what determines exposure? It could be an individual's decision to enter one state or another, an outside actor's decision, a planned random allocation carried out by an investigator, or some combination of these alternatives. The random variable D takes on values of $d_i = 1$ for each individual i who is an observed member of the treatment group and $d_i = 0$ for each individual i who is an observed member of the control group.

The *observed outcome* variable Y is then

$$Y = \begin{cases} Y^1, D = 1, \\ Y^0, D = 0, \end{cases}$$

or (this is called the *switching equation*)

$$Y = DY^1 + (1 - D)Y^0.$$

Notice what this definition really says. One can never observe the potential outcome under the treatment state for those observed in the control state, and one can never observe the potential outcome under the control state for those observed in the treatment state. This impossibility implies that one typically cannot calculate individual-level causal effects. Holland (1986) describes this challenge as *the fundamental problem of causal inference*.

Group/Potential Outcome	Y^1	Y^0
Treatment $D = 1$	Observable as Y	Counterfactual
Control $D = 0$	Counterfactual	Observable as Y

Treatment effects

The fundamental problem of causal inference does not mean, though, that we cannot estimate some average effect, though. In particular, we will often talk about the *average treatment effect* (ATE) defined as

$$E[\delta] \equiv E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$$

In general, the average treatment effect is equal to the expected value of the what-if difference in outcomes for a randomly selected individual from the population.

> **Note**
>
> The expectation is defined with reference to the population of interest. In the training example, the population would be defined as "all adults eligible for training," and eligibility would need to be defined carefully. Thus, to define average causal effects and then interpret estimates of them, one must clearly define the characteristics of the individuals in the assumed population of interest.

While the ATE is often quantity of interest, sometimes we are interested in other treatment effects. One can also define conditional ATEs based on the exposure status. The average treatment effect for the treated (ATT) is

$$E[\delta \mid D = 1] = E[Y^1 - Y^0 \mid D = 1] = E[Y^1 \mid D = 1] - E[Y^0 \mid D = 1].$$

The average treatment effect for the untreated (ATU, sometimes also denoted as ATC or ATUT) is

$$E[\delta \mid D = 0] = E[Y^1 - Y^0 \mid D = 0] = E[Y^1 \mid D = 0] - E[Y^0 \mid D = 0].$$

If we denote the probability of exposure to the treatment as $\pi \equiv P[D = 1]$, then the relationship between the ATE, ATT, and ATU is

$$\begin{aligned} ATE = E[\delta] &= E[\delta \mid D = 1]P[D = 1] + E[\delta \mid D = 0]P[D = 0] \\ &= \pi ATT + (1 - \pi)ATU \\ &= \pi E[Y^1 \mid D = 1] + (1 - \pi)E[Y^1 \mid D = 0] \\ &- (\pi E[Y^0 \mid D = 1] + (1 - \pi)E[Y^0 \mid D = 0]) \end{aligned}$$

> **Note**
>
> For the estimation of the ATE we need five quantities, out of which two are not observed (counterfactual outcomes).

What is the conceptual difference between all these effects? Recall our opening example. The ATT was the Gryffindor effect for those who were assigned to Gryffindor by the Sorting Hat. It was positive. The ATU was the Gryffindor effect for those who were assigned to Slytherin by the Sorting Hat. It was negative and equal in the absolute value. Since half of the students went to Gryffindor, the ATE was exactly zero.

Consider the college example. The ATT is the average effect of college on income of those who decided to go to college, rather than across all students who could potentially attend college. The difference between the ATE and ATT can also be understood with reference to individuals. From this perspective, the ATE is the expected what-if difference in income that would be observed if we could assign a *randomly selected individual* to both college and no-college treatments. In contrast, the ATT is the expected what-if difference in income that would be observed if we could assign a *randomly selected college student* to both college and no-college treatments.

For this example, the ATT is a theoretically important quantity, for if there is no college effect for college students, then it is unlikely that there would be a college effect for students who typically do not go to college. And, if policy interest were focused on whether college is beneficial for college students (and thus whether public support of colleges is a reasonable government expenditure, etc.), then the college effect for college students is the only quantity we would want to estimate. The ATU would be of interest as well if the goal of analysis is ultimately to determine the effect of a potential policy intervention, such as a college voucher program, designed to move more students to college.

Naive treatment effect

What happens if we go ahead and try to estimate the Naive Treatment Effect (NTE) using the data that we observe? The NTE is

$$NTE \equiv E[Y^1 \mid D = 1] - E[Y^0 \mid D = 0].$$

This is the effect that we can estimate from the data, since it does not involve any counterfactuals. But does it correspond to any of the treatment effects of interest that we introduced previously?

✓ Math Time: Decomposition of the NTE

We start with the definition of the NTE

$$\begin{aligned} NTE &= E[Y^1 \mid D=1] - E[Y^0 \mid D=0] \\ &= \underbrace{E[Y^1 \mid D=1] - E[Y^0 \mid D=1]}_{ATT} + E[Y^0 \mid D=1] - E[Y^0 \mid D=0] \end{aligned}$$

Now let's rewrite the ATT as

$$\begin{aligned} ATT &= \pi ATT + (1-\pi)ATT \\ &= \underbrace{\pi ATT + (1-\pi)ATU}_{ATE} - (1-\pi)ATU + (1-\pi)ATT \\ &= ATE + (1-\pi)(ATT - ATU). \end{aligned}$$

Finally, we can re-write the NTE as

$$NTE = ATE + \underbrace{E[Y^0 \mid D=1] - E[Y^0 \mid D=0]}_{\text{selection bias}} + \underbrace{(1-\pi)(ATT - ATU)}_{\text{diff treatment eff bias}}$$

We can arrive at an alternative decomposition that starts with ATU instead of ATT.

$$NTE = ATE + E[Y^1 \mid D=1] - E[Y^1 \mid D=0] + \pi(ATU - ATT).$$

? Homework

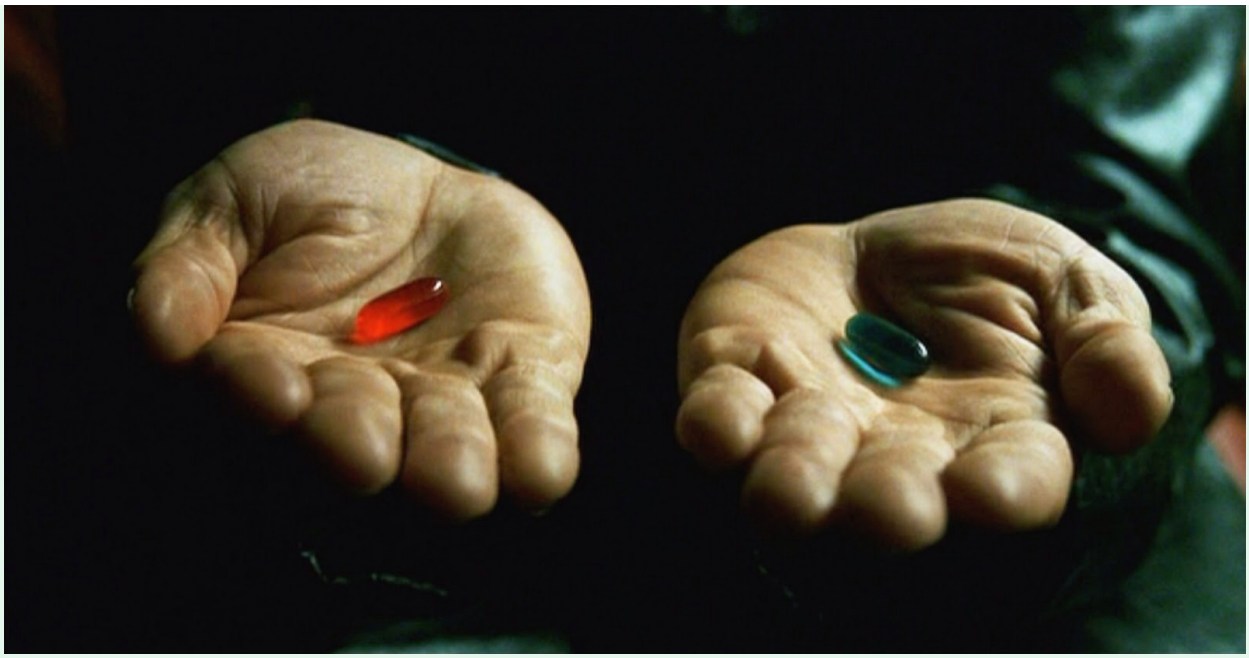
Prove that

$$NTE = ATE + E[Y^1 \mid D=1] - E[Y^1 \mid D=0] + \pi(ATU - ATT)$$

Thus, we see that the NTE is decomposed into the ATE, the difference between potential outcomes (either Y^0 or Y^1) among the treatment and control groups (*selection bias*), and the difference between ATT and ATU (*differential treatment effect bias*). So in general, unless no biases are present, the NTE will not be equal to the ATE.

Let's work through a few examples.

✓ Math Time: Example (The Matrix)



Group	$E[Y^1 \mid D]$	$E[Y^0 \mid D]$	$E[\delta \mid D]$	$P(D)$
Red Pill ($D = 1$)	8	2	6	0.2
Blue Pill ($D = 0$)	3	5	-2	0.8

The NTE is $8 - 5 = 3$. The ATE is $0.2 \times 6 - 0.8 \times 2 = 1.2 - 1.6 = -0.4$. Then we can decompose the NTE as

$$NTE = -0.4 + (2 - 5) + 0.8(6 + 2) = -0.4 - 3 + 6.4 = 3$$

or

$$NTE = -0.4 + (8 - 3) + 0.2(-2 - 6) = -0.4 + 5 - 1.6 = 3$$

✓ **Math Time: Example (The Sorting Hat)**

Group	$E[Y^1 \mid D]$	$E[Y^0 \mid D]$	$E[\delta \mid D]$	$P(D)$
Gryffindor ($D = 1$)	38.6	21	17.6	0.5
Slytherin ($D = 0$)	13.4	31	-17.6	0.5

The NTE is $38.6 - 31 = 7.6$. The ATE is $0.5 \times 17.6 - 0.5 \times 17.6 = 0$. Then we can decompose the NTE as

$$NTE = 0 + (21 - 31) + 0.5(17.6 + 17.6) = -10 + 17.6 = 7.6$$

or

$$NTE = 0 + (38.6 - 13.4) + 0.5(-17.6 - 17.6) = 25.2 - 17.6 = 7.6$$

Treatment assignment and randomization

We have seen that in the observational studies, when individuals are not randomly assigned to treatment groups, the naive way of estimating the ATE does not in general give us the ATE. The NTE is biased. But let's go back to the experimental ideal and see where the magic happens.

The main feature of a randomized experiment is that treatment assignment is done randomly, independent of potential outcomes. The formal way of writing this independence feature is

$$(Y^0, Y^1) \perp D.$$

D is jointly independent of all functions of the potential outcomes. This assumption implies that

$$E[Y^1 \mid D = 1] = E[Y^1 \mid D = 0]$$
$$E[Y^0 \mid D = 1] = E[Y^0 \mid D = 0]$$

These equalities automatically kill the selection bias.

How about the differential treatment effect bias? Notice that

$$ATT = E[Y^1 \mid D = 1] - E[Y^0 \mid D = 1] = E[Y^1 \mid D = 0] - E[Y^0 \mid D = 0] = ATU.$$

This, in turn, implies that

$$ATE = \pi ATT + (1 - \pi) ATU = ATT = ATU.$$

Under independence, both conditional treatment effects are identical and equal to the ATE. Hence, independence also kills that differential treatment effect bias. You can see then that our NTE becomes, in fact, the ATE. Thus, in a randomized experiment a simple difference in mean outcomes in the treatment and control groups is an unbiased estimator of the ATE.

Note

The independence assumption does not imply that $E[Y^1 \mid D = 1] = E[Y^0 \mid D = 1]$ or that $E[Y^1 \mid D = 0] = E[Y^0 \mid D = 0]$. In other words, exposure is independent of potential outcomes but observed outcomes and exposure can be related.

In the observational data, the independence assumption typically does not hold.

> **Rosenbaum (2002)**
>
> An observational study is an empiric investigation of treatments, policies, or exposures and the effects they cause, but it differs from an experiment in that the investigator cannot control the assignment of treatments to subjects.

The first step in the analysis of observational data is to investigate the *treatment selection mechanism*. Notice the switch in language from *assignment* to *selection*. And this is challenging! Although some processes by which individuals select alternative treatments can be examined empirically, a full accounting of treatment selection is sometimes impossible (e.g., if subjects are motivated to select on the causal effect itself and a researcher does not have a valid measure of their expectations).

While the independence assumptions will almost surely not hold in the observational data, some other assumptions can be more defendable. For example, consider the following two assumptions:

$$\begin{aligned}\text{Assumption 1: } & E[Y^1 \mid D = 1] = E[Y^1 \mid D = 0] \\ \text{Assumption 2: } & E[Y^0 \mid D = 1] = E[Y^0 \mid D = 0]\end{aligned}$$

Can the naive estimate say anything useful about our treatment effects of interest if at least of these assumptions holds?

> ✓ **Math Time**
>
> If only Assumption 1 holds, then the decomposition implies that
>
> $$\begin{aligned} NTE &= E[Y^1 \mid D = 1] - E[Y^0 \mid D = 0] \\ &= E[Y^1 \mid D = 1] + E[Y^1 \mid D = 0] - E[Y^1 \mid D = 0] - E[Y^0 \mid D = 0] \\ &= ATU + E[Y^1 \mid D = 1] - E[Y^1 \mid D = 0] \\ &= ATU. \end{aligned}$$
>
> If only Assumption 2 holds, then the decomposition implies that
>
> $$\begin{aligned} NTE &= E[Y^1 \mid D = 1] - E[Y^0 \mid D = 0] \\ &= E[Y^1 \mid D = 1] + E[Y^0 \mid D = 1] - E[Y^0 \mid D = 1] - E[Y^0 \mid D = 0] \\ &= ATT + E[Y^0 \mid D = 1] - E[Y^0 \mid D = 0] \\ &= ATT. \end{aligned}$$

These possibilities can be important in practice. For some applications, it may be the case that we have good theoretical reason to believe that

- Assumption 2 is valid because those in the treatment group would, on average, do no better or no worse under the control than those in the control group
- Assumption 1 is invalid because those in the control group would not do nearly as well under the treatment as those in the treatment group.

Under this scenario, the naive estimator will deliver an unbiased and consistent estimate of the ATT, even though it is still biased and inconsistent for both the ATU and the ATE.

The path forward in the analysis of the observation data is to try to condition or stratify the data or find some features of the data that allow us to defend either one of these assumptions or both of them.

Regression Interpretation

Let's write the switching equation as

$$\begin{aligned} Y &= Y^0 + (Y^1 - Y^0)D \\ &= E[Y^0] + (Y^1 - Y^0)D + Y^0 - E[Y^0]. \end{aligned}$$

Suppose that the individual treatment effects are the same for everyone, so that $\delta = Y^1 - Y^0$ is a constant. Note that this implies that $ATE = ATT = ATU = \delta$.
Now denote $\alpha \equiv E[Y^0]$ and $\epsilon \equiv Y^0 - E[Y^0]$. The switching equation becomes

$$Y = \alpha + \delta D + \epsilon.$$

This looks like a regression equation, where α is the constant, δ is the coefficient of interest, D is the independent variable, and ϵ is the error term.

Consider the conditional expectations of Y under different values of D.

$$E[Y \mid D = 1] = \alpha + \delta + E[\epsilon \mid D = 1]$$

and

$$E[Y \mid D = 0] = \alpha + E[\epsilon \mid D = 0].$$

Then the difference between the two is

$$E[Y \mid D = 1] - E[Y \mid D = 0] = \delta + E[\epsilon \mid D = 1] - E[\epsilon \mid D = 0].$$

As long as D is independent of the error term, the difference on the left identifies the ATE (=ATT=ATU). However, if the error term and D are correlated, the estimate of δ will be biased because of the selection bias (there is no differential treatment effect bias in this case).

SUTVA

The potential outcomes framework relies on a simple but strong assumption known as the stable unit treatment value assumption or *SUTVA* (Rubin 1980b, 1986). In economics, this is sometimes referred to as a no-macro-effect or partial equilibrium assumption (Heckman 2000, 2005). SUTVA is a basic assumption of causal effect stability that requires that the potential outcomes of individuals be unaffected by the treatment exposures of other individuals.

> **Rubin (1986)**
>
> SUTVA is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive.

Generally, suppose that d is an $N \times 1$ vector of treatment indicator variables for N individuals, and define potential outcomes generally as functions of the vector d. The outcome for individual i under the treatment is $y_i^1(d)$, and the outcome for individual i under the control is $y_i^0(d)$. Accordingly, the individual-level causal effect for individual i is $\delta(d)$. SUTVA is what allows us to write $y_i^1(d) = y_i^1$ and $y_i^0(d) = y_i^0$ and, as a result, assert that individual-level causal effects $\delta_i = \delta_i(d)$ exist that are independent of the assignment process.

SUTVA implies three things:

- each individual within a treatment group receives the same treatment (e.g., all colleges provide a similar education, there are no good or bad colleges)
- no spillovers to other individuals' potential outcomes when an individual is exposed to some treatment (e.g., your neighbor who goes to college does not affect your outcomes if you do not go to college)
- no general equilibrium effects (e.g., we are not comparing a case when no one goes to college vs. when everyone goes to college, because that would lead to dramatic changes in the labor market)

When does SUTVA break down? The most typical example is that of peer effects: your peer's exposure to the treatment might affect your potential outcomes. Suppose we conduct a randomized experiment on the effect of a new textbook on academic performance in high school. We randomly assign different classes to the treatment group (receive the new textbook) and the control group (use the old textbook). If students from the treatment group interact with the students from the control group and share the new textbook, then we might find a zero treatment effect of the new textbook. However, SUTVA is violated here.

Consider the college example. For SUTVA to hold, the college effect cannot be a function of the number (and/or composition) of students who enroll into college. For a variety of reasons—endogenous peer effects, capacity constraints, and so on—most

researchers would probably expect that the college effect would change if large numbers of non-college educated students entered college. There may be good theoretical reasons to believe that macro effects would emerge if the college enrollment ballooned. As a result, it may be that researchers can estimate the causal effect of college only for those who would typically choose to attend college. But the effect would be subject to the constraint that the proportion of students in college remain relatively constant. Accordingly, it may be impossible to determine from any data that could be collected what the college effect on income would be under a new distribution of students that would result from a large and effective policy intervention.

In general, if SUTVA is maintained, but there is some doubt about its validity, then certain types of marginal effect estimates can usually still be defended. The idea here would be to state that the estimates of average causal effects hold only for what-if movements of a very small number of individuals from one hypothetical treatment state to another. If more extensive what-if contrasts are of interest, such as would be induced by a widespread intervention, then SUTVA would need to be dropped and variation of the causal effect as a function of treatment assignment patterns would need to be modeled explicitly.