is at least as sensible as a ratio metric for all of the examples we will consider. (The ratio metric might be preferable if we were attempting to make effect comparisons across outcomes with very different base rates, such as the effect of the same treatment on pancreatic cancer and hypertension.)

## 2.5 The Stable Unit Treatment Value Assumption

In most applications, the potential outcome model retains its tractability through the maintenance of a strong assumption known as the stable unit treatment value assumption or SUTVA (see Rubin 1980b, 1986). In economics, a version of this assumption is sometimes referred to as a no-macro-effect or partial equilibrium assumption (see Garfinkel, Manski, and Michalopoulos 1992, Heckman 2000, 2005, for the history of these ideas, and Manski and Garfinkel 1992 for examples).[13]

SUTVA, as implied by its name, is a basic assumption of causal effect stability that requires that the potential outcomes of individuals be unaffected by changes in the treatment exposures of all other individuals. In the words of Rubin (1986:961), who developed the term,

> 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.

Consider the idealized example in Table 2.2, in which SUTVA is violated because the treatment effect varies with treatment assignment patterns. For the idealized example, there are three randomly drawn subjects from a population of interest, and the study is designed such that at least one of the three study subjects must receive the treatment and at least one must receive the control. The first column of the table gives the six possible treatment assignment patterns.[14] The first row of Table 2.2 presents all three ways to assign one individual to the treatment and the other two to the control, as well as the potential outcomes for each of the three subjects. Subtraction within the last column shows that the individual-level causal effect is 2 for all three individuals. The second row of Table 2.2 presents all three ways to assign two individuals to the treatment and one to the control. As shown in the last column of the row, the individual-level causal effects implied by the potential outcomes are now 1 instead of 2. Thus, for this idealized example, the underlying causal effects are a function of the treatment assignment patterns, such that the treatment is less effective when more individuals are assigned to it. For SUTVA to hold, the potential outcomes would need to be identical for both rows of the table.

[13] SUTVA is a much maligned acronym, and many others use different labels. Manski (2013a:S1), for example, has recently labeled the same assumption the "individualistic treatment response" assumption in order "to mark it as an assumption that restricts the form of treatment response functions."

[14] For this example, assume that the values of $y_i^1$ and $y_i^0$ for each individual $i$ are either deterministic potential outcomes or exactly equal to $E[Y_i^1]$ and $E[Y_i^0]$ for each individual $i$. Also, assume that these three subjects comprise a perfectly representative sample of the population.

**Table 2.2** A Hypothetical Example in Which SUTVA Is Violated

| Treatment assignment patterns | Potential outcomes |
|---|---|
| $\begin{bmatrix} d_1=1 \\ d_2=0 \\ d_3=0 \end{bmatrix}$ or $\begin{bmatrix} d_1=0 \\ d_2=1 \\ d_3=0 \end{bmatrix}$ or $\begin{bmatrix} d_1=0 \\ d_2=0 \\ d_3=1 \end{bmatrix}$ | $y_1^1=3$ $y_1^0=1$<br>$y_2^1=3$ $y_2^0=1$<br>$y_3^1=3$ $y_3^0=1$ |
| $\begin{bmatrix} d_1=1 \\ d_2=1 \\ d_3=0 \end{bmatrix}$ or $\begin{bmatrix} d_1=0 \\ d_2=1 \\ d_3=1 \end{bmatrix}$ or $\begin{bmatrix} d_1=1 \\ d_2=0 \\ d_3=1 \end{bmatrix}$ | $y_1^1=2$ $y_1^0=1$<br>$y_2^1=2$ $y_2^0=1$<br>$y_3^1=2$ $y_3^0=1$ |

This type of treatment effect dilution is only one way in which SUTVA can be violated. More generally, suppose that $\mathbf{d}$ is an $N \times 1$ vector of treatment indicator variables for $N$ individuals (analogous to the treatment assignment vectors in the first column of Table 2.2), and define potential outcomes of each individual as functions across all potential configurations of the elements of vector $\mathbf{d}$. Accordingly, the outcome for individual $i$ under the treatment is $y_i^1(\mathbf{d})$, and the outcome for individual $i$ under the control is $y_i^0(\mathbf{d})$. The treatment effect for each individual $i$ is then

$$\delta_i(\mathbf{d}) = y_i^1(\mathbf{d}) - y_i^0(\mathbf{d}). \tag{2.5}$$

With this more general setup, individual-level treatment effects could be different for every possible pattern of treatment exposure.

SUTVA is what allows us to declare $y_i^1(\mathbf{d}) = y_i^1$ and $y_i^0(\mathbf{d}) = y_i^0$ and, as a result, assert that individual-level causal effects $\delta_i$ exist that are independent of the overall configuration of causal exposure. If SUTVA cannot be maintained, then the simplified definition in Equation (2.1) is invalid, and the individual-level treatment effect must be written in its most general form in Equation (2.5), with all ensuing analysis proceeding conditional on alternative vectors $\mathbf{d}$.

Sometimes it is argued that SUTVA is so restrictive that we need an alternative conception of causality for the social sciences. Our position is that SUTVA reveals the limitations of social science data and the perils of immodest causal modeling rather than the limitations of the potential outcome model itself. Rather than consider SUTVA as overly restrictive, researchers should always reflect on the plausibility of SUTVA in each application and use such reflection to motivate a clear discussion of the meaning and scope of all causal effect estimates offered. Such reflection may lead one to determine that only the more general case of the potential outcome framework can be justified, and this may necessitate building the analysis on top of the individual-level treatment effect defined in Equation (2.5) rather than the SUTVA-simplified variant in Equation (2.1). In some cases, however, analysis can proceed assuming SUTVA, as long as all resulting estimates are given restricted interpretations, as we now explain.

Typical SUTVA violations share two interrelated features: (1) influence patterns that result from contact across individuals in social or physical space and (2) dilution/concentration patterns that one can assume would result from changes in the prevalence of the treatment. Neither feature is entirely distinct from the other, and in many cases dilution/concentration effects arise because influence patterns are present. Yet, if the violation can be interpreted as a dilution/concentration pattern, even when generated in part by an underlying influence pattern, then the analyst can proceed by scaling back the asserted relevance of any estimates to situations where the prevalence of the treatment is not substantially different.

For a simple example, consider the worker training example. Here, the plausibility of SUTVA may depend on the particular training program. For small training programs situated in large labor markets, the structure of wage offers to retrained workers may be entirely unaffected by the existence of the training program. However, for a sizable training program in a small labor market, it is possible that the wages on offer to retrained workers would be a function of the way in which the price of labor in the local labor market responds to the movement of trainees in and out of the program (as might be the case in a small company town after the company has just gone out of business and a training program is established). As a result, SUTVA may be reasonable only for a subset of the training sites for which data have been collected.

For an example of where influence patterns are more of a threat to SUTVA, consider the example of the Catholic school effect. For SUTVA to hold, the effectiveness of Catholic schooling cannot be a function of the number (and/or composition) of students who enter the Catholic school sector. For a variety of reasons – endogenous peer effects, capacity constraints, and so on – most school effects researchers would probably expect that the Catholic school effect would change if large numbers of public school students entered the Catholic school sector. As a result, because there are good theoretical reasons to believe that the pattern of effects would change if Catholic school enrollments ballooned, it may be that researchers can estimate the causal effect of Catholic schooling only for those who would typically choose to attend Catholic schools, but also subject to the constraint that the proportion of students educated in Catholic schools remains constant. Accordingly, it may be impossible to determine from any data that could be collected what the Catholic school effect on achievement would be under a new distribution of students across school sectors that would result from a large and effective policy intervention. As a result, the implications of research on the Catholic school effect for research on school voucher programs may be quite limited, and this has not been clearly enough recognized by some (see Howell and Peterson 2002, chapter 6). A similar argument applies to research on charter school effects.

Consider a SUTVA violation for a related example: the evaluation of the effectiveness of mandatory school desegregation plans in the 1970s on the subsequent achievement of black students. Gathering together the results of a decade of research, Crain and Mahard (1983) conducted a meta-analysis of 93 studies of the desegregation effect on achievement. They argued that the evidence suggests an increase of .3 standard

deviations in the test scores of black students across all studies.[15] It seems undeniable that SUTVA is violated for this example, as the effect of moving from one school to another must be a function of relative shifts in racial composition across schools. Breaking the analysis into subsets of cities where the compositional shifts were similar could yield conditional average treatment effect estimates that can be more clearly interpreted. In this case, SUTVA would be abandoned in the collection of all desegregation events, but it could then be maintained for some groups (perhaps in cities where the compositional shift was comparatively small).

In general, if SUTVA is maintained but there is some doubt about its validity because dilution or concentration patterns would emerge under shifts in treatment prevalence, then certain types of marginal effect estimates can usually still be defended. The idea here is to state that the estimates of average causal effects hold only for what-if movements of relatively small numbers of individuals from one hypothetical treatment state to another.

If, however, influence patterns are inherent to the causal process of interest, and the SUTVA violation cannot be considered as a type of dilution or concentration, then it will generally not be possible to circumvent the SUTVA violation by proceeding with the same analysis and only offering cautious and conditional interpretations. The most well-developed literature on situations such as these is the literature on the effects of vaccine programs (see Hudgens and Halloran 2008). Here, additional causal effects of interest using the potential outcome framework have been defined, conditional on the overall pattern of treatment assignment:

> The *indirect effect* of a vaccination program or strategy on an individual is the difference between what the outcome is in the individual not being vaccinated in a community with the vaccination program and what the outcome would have been in the individual, again not being vaccinated, but in a comparable community with no vaccination program. It is, then, the effect of the vaccination program on an individual who was not vaccinated. The combined *total effect* in an individual of being vaccinated and the vaccination program in the community is the difference between the outcome in the individual being vaccinated in a community with the vaccination program and what the outcome would be if the individual were not vaccinated and the community did not have the vaccination program. The total effect, then, is the effect of the vaccination program combined with the effect of the person having been vaccinated. The *overall effect* of a vaccination program is the difference in the outcome in an average individual

[15] As reviewed by Schofield (1995) and noted in Clotfelter (2004), most scholars now accept that the evidence suggests that black students who were bused to predominantly white schools experienced small positive reading gains but no substantial mathematics gains. Cook and Evans (2000:792) conclude that "it is unlikely that efforts at integrating schools have been an important part of the convergence in academic performance [between whites and blacks], at least since the early 1970s" (see also Armor 1995; Rossell, Armor, and Walberg 2002). Even so, others have argued that the focus on test score gains has obscured some of the true effectiveness of desegregation. In a review of these longer-term effects, Wells and Crain (1994:552) conclude that "interracial contact in elementary and secondary school can help blacks overcome perpetual segregation."

> in a community with the vaccination program compared to an average individual in a comparable population with no vaccination program. (Halloran, Longini, and Struchiner 2010:272; italics in the original)

Effectively estimating these types of effects generally requires a nested randomization structure, wherein (1) vaccine programs are randomly assigned to a subset of participating groups and then (2) vaccinations are randomly given to individuals within groups enrolled in vaccine programs. These particular study designs are not possible for most social science applications, but the basic interpretive framework has been adopted to clarify what can be learned from social experiments, in particular, the Moving to Opportunity neighborhood experiment (see Sobel 2006).[16]

Much observational research on social influence patterns proceeds without consideration of these sorts of complications. Consider the contentious literature on whether peer effects have accelerated the obesity epidemic, as presented in Section 1.3 (see page 26). As we noted there, the basic claim of Christakis and Fowler (2007) is that having a friend who becomes obese increases one's own odds of becoming obese. Yet, their full set of claims is substantially more detailed, suggesting that these peer effects travel across network paths of length three before dying out. In particular, one's odds of becoming obese also increase if friends of friends become obese and if friends of friends of friends become obese. The sizes of these three effects diminish with the length of friendship distance.

Now consider whether SUTVA is reasonable for such a schedule of effects across network ties. Holding the social network structure fixed, if obesity increases in the population, then, on average, individuals have more obese friends, more obese friends of friends, and more obese friends of friends of friends. Most theoretical predictions would suggest that the effects on one's own odds of becoming obese that result from having friends of friends of friends who become obese should decline with the proportion of one's own friends who are already obese or who have just become obese.[17] Effects that cascade in these conditional ways, because they are defined across a pattern of interpersonal contact between units, nearly always violate SUTVA.[18]

---

[16] Suitable models for observational data are an active frontier of research (see Hong and Raudenbush 2013; Manski 2013a). Tchetgen Tchetgen and VanderWeele (2010) show that some estimators may be effective for applications with observational data if all relevant patterns of treatment assignment (i.e., $\mathbf{d}$) can be attributed to measured treatment-level variables.

[17] This means that, even if the issues raised by critics on the severity of homophily bias are invalid (see VanderWeele 2011b for a convincing case that they have been exaggerated), the pattern of effects only holds under the prevalence of obesity in the data analyzed, which is the pattern of obesity in Framingham, Massachusetts, among adults born in 1948 for whom data was collected between 1971 and 1999 (and for a social network structure elicited by an unconventional name generator). The overall pattern of declining effects may be valid, but the relation of the various lagged regression coefficients offered to well-defined causal effects of general interest may be rather thin.

[18] When we have conveyed this point to network analysis researchers, a common reaction is that the potential outcome model must not, therefore, be suitable for studying causal effects that propagate across networks. The logic of this position eludes us for two reasons. First, the potential outcome model cannot be deemed inappropriate because it makes clear how hard it is to define and estimate the effects that analysts claim that they wish to estimate. Second, the potential outcome model can accommodate SUTVA violations, although not without considerable additional effort. Weihua An (2013) demonstrates the value of counterfactual thinking for modeling peer effects, fully embedded within a social network perspective (see also VanderWeele and An 2013).