Chapter 5
INTERACTION

Consider again a randomized experiment to answer the causal question "does one's looking up at the sky make other pedestrians look up too?" We have so far restricted our interest to the causal effect of a single treatment (looking up) in either the entire population or a subset of it. However, many causal questions are actually about the effects of two or more simultaneous treatments. For example, suppose that, besides randomly assigning your looking up, we also randomly assign whether you stand in the street dressed or naked. We can now ask questions like: what is the causal effect of your looking up if you are dressed? And if you are naked? If these two causal effects differ we say that the two treatments under consideration (looking up and being dressed) interact in bringing about the outcome.

When joint interventions on two or more treatments are feasible, the identification of interaction allows one to implement the most effective interventions. Thus understanding the concept of interaction is key for causal inference. This chapter provides a formal definition of interaction between two treatments, both within our already familiar counterfactual framework and within the sufficient-component-cause framework.

5.1 Interaction requires a joint intervention

Suppose that in our heart transplant example, individuals were assigned to receiving either a multivitamin complex ($E = 1$) or no vitamins ($E = 0$) before being assigned to either heart transplant ($A = 1$) or no heart transplant ($A = 0$). We can now classify all individuals into 4 treatment groups: vitamins-transplant ($E = 1$, $A = 1$), vitamins-no transplant ($E = 1$, $A = 0$), no vitamins-transplant ($E = 0$, $A = 1$), and no vitamins-no transplant ($E = 0$, $A = 0$). For each individual, we can now imagine 4 potential or counterfactual outcomes, one under each of these 4 treatment combinations: $Y^{a=1,e=1}$, $Y^{a=1,e=0}$, $Y^{a=0,e=1}$, and $Y^{a=0,e=0}$. In general, an individual's counterfactual outcome $Y^{a,e}$ is the outcome that would have been observed if we had intervened to set the individual's values of A and E to a and e, respectively. We refer to interventions on two or more treatments as *joint interventions*.

The counterfactual Y^a corresponding to an intervention on A alone is the joint counterfactual $Y^{a,e}$ if the observed E takes the value e, i.e., $Y^a = Y^{a,E}$. In fact, consistency is a special case of this recursive substitution. Specifically, the observed $Y = Y^A = Y^{A,E}$, which is our definition of consistency. See also Technical Point 6.2.

We are now ready to provide a definition of interaction within the counterfactual framework. There is interaction between two treatments A and E if the causal effect of A on Y after a joint intervention that set E to 1 differs from the causal effect of A on Y after a joint intervention that set E to 0. For example, there would be an interaction between transplant A and vitamins E if the causal effect of transplant on survival had everybody taken vitamins were different from the causal effect of transplant on survival had nobody taken vitamins.

When the causal effect is measured on the risk difference scale, we say that there is *interaction between A and E on the additive scale* in the population if

$$\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=1}=1\right] \neq \Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right].$$

For example, suppose the causal risk difference for transplant A when everybody receives vitamins, $\Pr\left[Y^{a=1,e=1}=1\right] - \Pr\left[Y^{a=0,e=1}=1\right]$, were 0.1,

and that the causal risk difference for transplant A when nobody receives vitamins, $\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$, were 0.2. We say that there is interaction between A and E on the additive scale because the risk difference $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=1}=1\right]$ is less than the risk difference $\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$. Using simple algebra, it can be easily shown that this inequality implies that the causal risk difference for vitamins E when everybody receives a transplant, $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=1,e=0}=1\right]$, is also less than the causal risk difference for vitamins E when nobody receives a transplant A, $\Pr\left[Y^{a=0,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$. That is, we can equivalently define interaction between A and E on the additive scale as

$$\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=1,e=0}=1\right]\neq\Pr\left[Y^{a=0,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right].$$

The two inequalities displayed above show that treatments A and E have equal status in the definition of interaction. See also Technical Point 5.1.

Let us now review the difference between interaction and effect modification. As described in the previous chapter, a variable V is a modifier of the effect of A on Y when the average causal effect of A on Y varies across levels of V. Note the concept of effect modification refers to the causal effect of A, not to the causal effect of V. For example, sex was an effect modifier for the effect of heart transplant in Table 4.1, but we never discussed the effect of sex on death. Thus, when we say that V modifies the effect of A we are not considering V and A as variables of equal status, because only A is considered to be a variable on which we could hypothetically intervene. That is, the definition of effect modification involves the counterfactual outcomes Y^a, not the counterfactual outcomes $Y^{a,v}$. In contrast, the definition of interaction between A and E gives equal status to both treatments A and E, as reflected by the two equivalent definitions of interaction shown above. The concept of interaction refers to the joint causal effect of two treatments A and E, and thus involves the counterfactual outcomes $Y^{a,e}$ under a joint intervention.

5.2 Identifying interaction

In previous chapters we have described the conditions that are required to identify the average causal effect of a treatment A on an outcome Y, either in the entire population or in a subset of it. The three key identifying conditions were exchangeability, positivity, and consistency. Because interaction is concerned with the joint effect of two (or more) treatments A and E, identifying interaction requires exchangeability, positivity, and consistency for both treatments.

Suppose that vitamins E were randomly, and unconditionally, assigned by the investigators. Then positivity and consistency hold, and the treated $E=1$ and the untreated $E=0$ are expected to be exchangeable. That is, the risk that would have been observed if all individuals had been assigned to transplant $A=1$ and vitamins $E=1$ equals the risk that would have been observed if all individuals who received $E=1$ had been assigned to transplant $A=1$. Formally, the marginal risk $\Pr\left[Y^{a=1,e=1}=1\right]$ is equal to the conditional risk $\Pr\left[Y^{a=1}=1|E=1\right]$. As a result, we can rewrite the definition of interaction between A and E on the additive scale as

$$\begin{aligned}&\Pr\left[Y^{a=1}=1|E=1\right]-\Pr\left[Y^{a=0}=1|E=1\right]\\&\neq\Pr\left[Y^{a=1}=1|E=0\right]-\Pr\left[Y^{a=0}=1|E=0\right],\end{aligned}$$

Technical Point 5.1

Interaction on the additive and multiplicative scales. The equality of causal risk differences $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=1}=1\right]=\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$ can be rewritten as

$$\Pr\left[Y^{a=1,e=1}=1\right]=\left\{\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}+\Pr\left[Y^{a=0,e=1}=1\right].$$

By subtracting $\Pr\left[Y^{a=0,e=0}=1\right]$ from both sides of the equation, we get $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]=$

$$\left\{\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}+\left\{\Pr\left[Y^{a=0,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}.$$

This equality is another compact way to show that treatments A and E have equal status in the definition of interaction.

When the above equality holds, we say that there is no *interaction between A and E on the additive scale*, and we say that the causal risk difference $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$ is additive because it can be written as the sum of the causal risk differences that measure the effect of A in the absence of E and the effect of E in the absence of A. Conversely, there is interaction between A and E on the additive scale if $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\neq$

$$\left\{\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}+\left\{\Pr\left[Y^{a=0,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}.$$

The interaction is *superadditive* if the 'not equal to' ($\neq$) symbol can be replaced by a 'greater than' ($>$) symbol. The interaction is *subadditive* if the 'not equal to' ($\neq$) symbol can be replaced by a 'less than' ($<$) symbol.

Analogously, one can define interaction on the multiplicative scale when the effect measure is the causal risk ratio, rather than the causal risk difference. We say that there is *interaction between A and E on the multiplicative scale* if

$$\frac{\Pr\left[Y^{a=1,e=1}=1\right]}{\Pr\left[Y^{a=0,e=0}=1\right]}\neq\frac{\Pr\left[Y^{a=1,e=0}=1\right]}{\Pr\left[Y^{a=0,e=0}=1\right]}\times\frac{\Pr\left[Y^{a=0,e=1}=1\right]}{\Pr\left[Y^{a=0,e=0}=1\right]}.$$

The interaction is *supermultiplicative* if the 'not equal to' ($\neq$) symbol can be replaced by a 'greater than' ($>$) symbol. The interaction is *submultiplicative* if the 'not equal to' ($\neq$) symbol can be replaced by a 'less than' ($<$) symbol.

which is exactly the definition of modification of the effect of A by E on the additive scale. In other words, when treatment E is randomly assigned, then the concepts of interaction and effect modification coincide. The methods described in Chapter 4 to identify modification of the effect of A by V can now be applied to identify interaction of A and E by simply replacing the effect modifier V by the treatment E.

Now suppose treatment E was not assigned by investigators. To assess the presence of interaction between A and E, one still needs to compute the four marginal risks $\Pr\left[Y^{a,e}=1\right]$. In the absence of marginal randomization, these risks can be computed for both treatments A and E, under the usual identifying assumptions, by standardization or IP weighting conditional on the measured covariates. An equivalent way of conceptualizing this problem follows: rather than viewing A and E as two distinct treatments with two possible levels (1 or 0) each, one can view AE as a combined treatment with four possible levels (11, 01, 10, 00). Under this conceptualization, the identification of interaction between two treatments is not different from the identification of the causal effect of one treatment that we have discussed in previous chapters. The same methods, under the same identifiability conditions, can be used. The only difference is that now there is a longer list of values that the treatment of interest can take, and therefore a greater number of counterfactual outcomes.

Sometimes one may be willing to assume (conditional) exchangeability for

treatment A but not for treatment E, e.g., when estimating the causal effect of A in subgroups defined by E in a randomized experiment. In that case, one cannot generally assess the presence of interaction between A and E, but can still assess the presence of effect modification by E. This is so because one does not need any identifying assumptions involving E to compute the effect of A in each of the strata defined by E. In the previous chapter we used the notation V (rather than E) for variables for which we are not willing to make assumptions about exchangeability, positivity, and consistency. For example, we concluded that the effect of transplant A was modified by nationality V, but we never required any identifying assumptions for the effect of V because we were not interested in using our data to compute the causal effect of V on Y. In Section 4.2 we argued on substantive grounds that V is a surrogate effect modifier; that is, V does not act on the outcome and therefore does not interact with A—no action, no interaction. But V is a modifier of the effect of A on Y because V is correlated with (e.g., it is a proxy for) an unidentified variable that actually has an effect on Y and interacts with A. Thus there can be modification of the effect of A by another variable without interaction between A and that variable.

Interaction between A and E without modification of the effect of A by E is also logically possible, though probably rare, because it requires dual effects of A and exact cancellations (VanderWeele 2009b).

In the above paragraphs we have argued that a sufficient condition for identifying interaction between two treatments A and E is that exchangeability, positivity, and consistency are all satisfied for the joint treatment (A, E) with the four possible values $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. Then standardization or IP weighting can be used to estimate the joint effects of the two treatments and thus to evaluate interaction between them. In Part III, we show that this condition is not necessary when the two treatments occur at different times. For the remainder of Part I (except this chapter) and most of Part II, we will focus on the causal effect of a single treatment A.

In Chapter 1 we described deterministic and nondeterministic counterfactual outcomes. Up to here, we used deterministic counterfactuals for simplicity. However, none of the results we have discussed for population causal effects and interactions require deterministic counterfactual outcomes. In contrast, the following section of this chapter only applies in the case that counterfactuals are deterministic. Further, we also assume that treatments and outcomes are dichotomous.

5.3 Counterfactual response types and interaction

Individuals can be classified in terms of their deterministic counterfactual responses. For example, in Table 4.1 (same as Table 1.1), there are four types of people: the "doomed" who will develop the outcome regardless of what treatment they receive (Artemis, Athena, Persephone, Ares), the "immune" who will not develop the outcome regardless of what treatment they receive (Demeter, Hestia, Hera, Hades), the "helped" who will develop the outcome only if untreated (Hebe, Kronos, Poseidon, Apollo, Hermes, Dyonisus), and the "hurt" who will develop the outcome only if treated (Rheia, Leto, Aphrodite, Zeus, Hephaestus, Polyphemus). Each combination of counterfactual responses is often referred to as a response pattern or a *response type*. Table 5.1 displays the four possible response types.

Table 5.1

Type	$Y^{a=0}$	$Y^{a=1}$
Doomed	1	1
Helped	1	0
Hurt	0	1
Immune	0	0

When considering two dichotomous treatments A and E, there are 16 possible response types because each individual has four counterfactual outcomes, one under each of the four possible joint interventions on treatments A and

E: $(1,1)$, $(0,1)$, $(1,0)$, and $(0,0)$. Table 5.2 shows the 16 response types for two treatments. This section explores the relation between response types and the presence of interaction in the case of two dichotomous treatments A and E and a dichotomous outcome Y.

Table 5.2

	$Y^{a,e}$ for each a,e value			
Type	$1,1$	$0,1$	$1,0$	$0,0$
1	1	1	1	1
2	1	1	1	0
3	1	1	0	1
4	1	1	0	0
5	1	0	1	1
6	1	0	1	0
7	1	0	0	1
8	1	0	0	0
9	0	1	1	1
10	0	1	1	0
11	0	1	0	1
12	0	1	0	0
13	0	0	1	1
14	0	0	1	0
15	0	0	0	1
16	0	0	0	0

Miettinen (1982) described the 16 possible response types under two binary treatments and outcome.

The first type in Table 5.2 has the counterfactual outcome $Y^{a=1,e=1}$ equal to 1, which means that an individual of this type would die if treated with both transplant and vitamins. The other three counterfactual outcomes are also equal to 1, i.e., $Y^{a=1,e=1} = Y^{a=0,e=1} = Y^{a=1,e=0} = Y^{a=0,e=0} = 1$, which means that an individual of this type would also die if treated with (no transplant, vitamins), (transplant, no vitamins), or (no transplant, no vitamins). In other words, neither treatment A nor treatment E has any effect on the outcome of such individual. He would die no matter what joint treatment he is assigned to. Now consider type 16. All the counterfactual outcomes are 0, i.e., $Y^{a=1,e=1} = Y^{a=0,e=1} = Y^{a=1,e=0} = Y^{a=0,e=0} = 0$. Again, neither treatment A nor treatment E has any effect on the outcome of an individual of this type. She would survive no matter what joint treatment she is assigned to. If all individuals in the population were of types 1 and 16, we would say that neither A nor E has any causal effect on Y; the sharp causal null hypothesis would be true for the joint treatment (A, E).

Let us now focus our attention on types 4, 6, 11, and 13. Individuals of type 4 would only die if treated with vitamins, whether they do or do not receive a transplant, i.e., $Y^{a=1,e=1} = Y^{a=0,e=1} = 1$ and $Y^{a=1,e=0} = Y^{a=0,e=0} = 0$. Individuals of type 13 would only die if not treated with vitamins, whether they do or do not receive a transplant, i.e., $Y^{a=1,e=1} = Y^{a=0,e=1} = 0$ and $Y^{a=1,e=0} = Y^{a=0,e=0} = 1$. Individuals of type 6 would only die if treated with transplant, whether they do or do not receive vitamins, i.e., $Y^{a=1,e=1} = Y^{a=1,e=0} = 1$ and $Y^{a=0,e=1} = Y^{a=0,e=0} = 0$. Individuals of type 11 would only die if not treated with transplant, whether they do or do not receive vitamins, i.e., $Y^{a=1,e=1} = Y^{a=1,e=0} = 0$ and $Y^{a=0,e=1} = Y^{a=0,e=0} = 1$.

Of the 16 possible response types in Table 5.2, we have identified 6 types (numbers 1, 4, 6, 11, 13, 16) with a common characteristic: for an individual with one of those response types, the causal effect of treatment A on the outcome Y is the same regardless of the value of treatment E, and the causal effect of treatment E on the outcome Y is the same regardless of the value of treatment A. In a population in which every individual has one of these 6 response types, the causal effect of treatment A in the presence of treatment E, as measured by the causal risk difference $\Pr\left[Y^{a=1,e=1}=1\right] - \Pr\left[Y^{a=0,e=1}=1\right]$, would equal the causal effect of treatment A in the absence of treatment E, as measured by the causal risk difference $\Pr\left[Y^{a=1,e=0}=1\right] - \Pr\left[Y^{a=0,e=0}=1\right]$. That is, if all individuals in the population have response types 1, 4, 6, 11, 13 and 16 then there will be no interaction between A and E on the additive scale.

Greenland and Poole (1988) noted that Miettinen's response types were not invariant to recoding of A and E (i.e., switching the labels "0" and "1"). They partitioned the 16 response types of Table 5.2 into these three equivalence classes that are invariant to recoding.

The presence of additive interaction between A and E implies that, for some individuals in the population, the value of their two counterfactual outcomes under $A = a$ cannot be determined without knowledge of the value of E, and vice versa. That is, there must be individuals in at least one of the following three classes:

1. those who would develop the outcome under only one of the four treatment combinations (types 8, 12, 14, and 15 in Table 5.2)
2. those who would develop the outcome under two treatment combinations, with the particularity that the effect of each treatment is exactly the opposite under each level of the other treatment (types 7 and 10)

Technical Point 5.2

Monotonicity of causal effects. Consider a setting with a dichotomous treatment A and outcome Y. The value of the counterfactual outcome $Y^{a=0}$ is greater than that of $Y^{a=1}$ only among individuals of the "helped" type. For the other 3 types, $Y^{a=1} \geq Y^{a=0}$ or, equivalently, an individual's counterfactual outcomes are monotonically increasing (i.e., nondecreasing) in a. Thus, when the treatment cannot prevent any individual's outcome (i.e., in the absence of "helped" individuals), all individuals' counterfactual response types are monotonically increasing in a. We then simply say that the causal effect of A on Y is monotonic.

The concept of monotonicity can be generalized to two treatments A and E. The causal effects of A and E on Y are monotonic if every individual's counterfactual outcomes $Y^{a,e}$ are monotonically increasing in both a and e. That is, if there are no individuals with response types $\left(Y^{a=1,e=1}=0, Y^{a=0,e=1}=1\right)$, $\left(Y^{a=1,e=1}=0, Y^{a=1,e=0}=1\right)$, $\left(Y^{a=1,e=0}=0, Y^{a=0,e=0}=1\right)$, and $\left(Y^{a=0,e=1}=0, Y^{a=0,e=0}=1\right)$.

3. those who would develop the outcome under three of the four treatment combinations (types 2, 3, 5, and 9)

On the other hand, the absence of additive interaction between A and E implies that either no individual in the population belongs to one of the three classes described above, or that there is a perfect cancellation of equal deviations from additivity of opposite sign. Such cancellation would occur, e.g., if there were an equal proportion of individuals of types 7 and 10, or of types 8 and 12.

For more on cancellations that result in additivity even when interaction types are present, see Greenland, Lash, and Rothman (2008).

The meaning of the term "interaction" is clarified by the classification of individuals according to their counterfactual response types (see also Fine Point 5.1). We now introduce a tool to conceptualize the causal mechanisms involved in the interaction between two treatments.

5.4 Sufficient causes

The meaning of interaction is clarified by the classification of individuals according to their counterfactual response types. We now introduce a tool to represent the causal mechanisms involved in the interaction between two treatments. Consider again our heart transplant example with a single treatment A. As reviewed in the previous section, some individuals die when they are treated, others when they are not treated, others die no matter what, and others do not die no matter what. This variety of response types indicates that treatment A is not the only variable that determines whether or not the outcome Y occurs.

Take those individuals who were actually treated. Only some of them died, which implies that treatment alone is insufficient to always bring about the outcome. As an oversimplified example, suppose that heart transplant $A=1$ only results in death in individuals allergic to anesthesia. We refer to the smallest set of background factors that, together with $A=1$, are sufficient to inevitably produce the outcome as U_1. The simultaneous presence of treatment $(A=1)$ and allergy to anesthesia $(U_1=1)$ is a minimal *sufficient cause* of the outcome Y.

Now take those individuals who were not treated. Again only some of them died, which implies that lack of treatment alone is insufficient to bring about the outcome. As an oversimplified example, suppose that no heart transplant

Fine Point 5.1

More on counterfactual types and interaction. The classification of individuals by counterfactual response types makes it easier to consider specific forms of interaction. For example, we may be interested in learning whether some individuals will develop the outcome when receiving both treatments $E = 1$ and $A = 1$, but not when receiving only one of the two. That is, whether individuals with counterfactual responses $Y^{a=1,e=1} = 1$ and $Y^{a=0,e=1} = Y^{a=1,e=0} = 0$ (types 7 and 8) exist in the population. VanderWeele and Robins (2007a, 2008) developed a theory of sufficient cause interaction for 2 and 3 treatments, and derived the identifying conditions for synergism that are described here. The following inequality is a sufficient condition for these individuals to exist:

$$\Pr\left[Y^{a=1,e=1} = 1\right] - \left(\Pr\left[Y^{a=0,e=1} = 1\right] + \Pr\left[Y^{a=1,e=0} = 1\right]\right) > 0$$

$$\text{or, equivalently, } \Pr\left[Y^{a=1,e=1} = 1\right] - \Pr\left[Y^{a=0,e=1} = 1\right] > \Pr\left[Y^{a=1,e=0} = 1\right]$$

That is, in an experiment in which treatments A and E are randomly assigned, one can compute the three counterfactual risks in the above inequality, and empirically check that individuals of types 7 and 8 exist.

Because the above inequality is a sufficient but not a necessary condition, it may not hold even if types 7 and 8 exist. In fact this sufficient condition is so strong that it may miss most cases in which these types exist. A weaker sufficient condition for synergism can be used if one knows, or is willing to assume, that receiving treatments A and E cannot prevent any individual from developing the outcome, i.e., if the effects are monotonic (see Technical Point 5.2). In this case, the inequality

$$\Pr\left[Y^{a=1,e=1} = 1\right] - \Pr\left[Y^{a=0,e=1} = 1\right] > \Pr\left[Y^{a=1,e=0} = 1\right] - \Pr\left[Y^{a=0,e=0} = 1\right]$$

is a sufficient condition for the existence of types 7 and 8. In other words, when the effects of A and E are monotonic, the presence of superadditive interaction implies the presence of type 8 (monotonicity rules out type 7). This sufficient condition for synergism under monotonic effects was originally reported by Greenland and Rothman in a previous edition of their book. It is now reported in Greenland, Lash, and Rothman (2008).

In genetic research it is sometimes interesting to determine whether there are individuals of type 8, a form of interaction referred to as *compositional epistasis*. VanderWeele (2010a) reviews empirical tests for compositional epistasis.

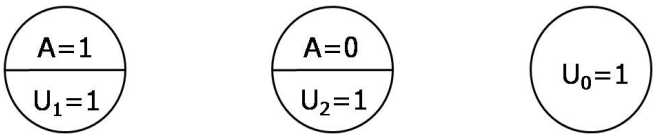
$A = 0$ only results in death if individuals have an ejection fraction less than 20%. We refer to the smallest set of background factors that, together with $A = 0$, are sufficient to produce the outcome as U_2. The simultaneous absence of treatment ($A = 0$) and presence of low ejection fraction ($U_2 = 1$) is another sufficient cause of the outcome Y.

Finally, suppose there are some individuals who have neither U_1 nor U_2 and that would have developed the outcome whether they had been treated or untreated. The existence of these "doomed" individuals implies that there are some other background factors that are themselves sufficient to bring about the outcome. As an oversimplified example, suppose that all individuals with pancreatic cancer at the start of the study will die. We refer to the smallest set of background factors that are sufficient to produce the outcome regardless of treatment status as U_0. The presence of pancreatic cancer ($U_0 = 1$) is another sufficient cause of the outcome Y.

By definition of background factors, the dichotomous variables U cannot be intervened on, and cannot be affected by treatment A.

We described 3 sufficient causes for the outcome: treatment $A = 1$ in the presence of U_1, no treatment $A = 0$ in the presence of U_2, and presence of U_0 regardless of treatment status. Each sufficient cause has one or more *components* $A = 1$ and $U_1 = 1$ in the first sufficient cause. Figure 5.1 represents each sufficient cause by a circle and its components as sections of the circle. The term *sufficient-component causes* is often used to refer to the sufficient causes and their components.

Figure 5.1



The graphical representation of sufficient-component causes helps visualize a key consequence of effect modification: as discussed in Chapter 4, the magnitude of the causal effect of treatment A depends on the distribution of effect modifiers. Imagine two hypothetical scenarios. In the first one, the population includes only 1% of individuals with $U_1 = 1$ (i.e., allergy to anesthesia). In the second one, the population includes 10% of individuals with $U_1 = 1$. The distribution of U_2 and U_0 is identical between these two populations. Now, separately in each population, we conduct a randomized experiment of heart transplant A in which half of the population is assigned to treatment $A = 1$. The average causal effect of heart transplant A on death will be greater in the second population because there are more individuals susceptible to develop the outcome if treated. One of the 3 sufficient causes, $A = 1$ plus $U_1 = 1$, is 10 times more common in the second population than in the first one, whereas the other two sufficient causes are equally frequent in both populations.

The graphical representation of sufficient-component causes also helps visualize an alternative concept of interaction, which is described in the next section. First we need to describe the sufficient causes for two treatments A and E. Consider our vitamins and heart transplant example. We have already described 3 sufficient causes of death: presence/absence of A (or E) is irrelevant, presence of transplant A regardless of vitamins E, and absence of transplant A regardless of vitamins E. In the case of two treatments we need to add 2 more ways to die: presence of vitamins E regardless of transplant A, and absence of vitamins regardless of transplant A. We also need to add four more sufficient causes to accommodate those who would die only under certain combination of values of the treatments A and E. Thus, depending on which background factors are present, there are 9 possible ways to die:

Greenland and Poole (1988) first enumerated these 9 sufficient causes.

1. by treatment A (treatment E is irrelevant)
2. by the absence of treatment A (treatment E is irrelevant)
3. by treatment E (treatment A is irrelevant)
4. by the absence of treatment E (treatment A is irrelevant)
5. by both treatments A and E
6. by treatment A and the absence of E
7. by treatment E and the absence of A
8. by the absence of both A and E
9. by other mechanisms (both treatments A and E are irrelevant)

In other words, there are 9 possible sufficient causes with treatment components $A = 1$ only, $A = 0$ only, $E = 1$ only, $E = 0$ only, $A = 1$ and $E = 1$, $A = 1$ and $E = 0$, $A = 0$ and $E = 1$, $A = 0$ and $E = 0$, and neither A nor E matter. Each of these sufficient causes includes a set of background factors from $U_1, \ldots, U_8$ and U_0. Figure 5.2 represents the 9 sufficient-component causes for two treatments A and E.

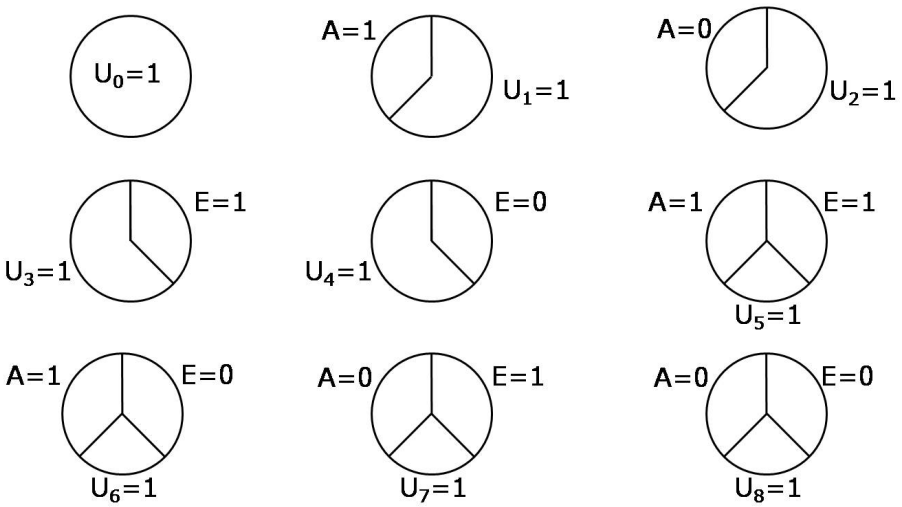


Figure 5.2

This graphical representation of sufficient-component causes is often referred to as "the causal pies."

Not all 9 sufficient-component causes for a dichotomous outcome and two treatments exist in all settings. For example, if receiving vitamins $E = 1$ does not kill any individual, regardless of her treatment A, then the 3 sufficient causes with the component $E = 1$ will not be present. The existence of those 3 sufficient causes would mean that some individuals (e.g., those with $U_3 = 1$) would be killed by receiving vitamins ($E = 1$), that is, their death would be prevented by not giving vitamins ($E = 0$) to them. See also Technical Point 5.3.

5.5 Sufficient cause interaction

The colloquial use of the term "interaction between treatments A and E" evokes the existence of some causal mechanism by which the two treatments work together (i.e., "interact") to produce certain outcome. Interestingly, the definition of interaction within the counterfactual framework does not require any knowledge about those mechanisms nor even that the treatments work together (see Fine Point 5.3). In our example of vitamins E and heart transplant A, we said that there is an interaction between the treatments A and E if the causal effect of A when everybody receives E is different from the causal effect of A when nobody receives E. That is, interaction is defined by the contrast of counterfactual quantities, and can therefore be identified by conducting an ideal randomized experiment in which the conditions of exchangeability, positivity, and consistency hold for both treatments A and E. There is no need to contemplate the causal mechanisms (physical, chemical, biologic, sociological...) that underlie the presence of interaction.

This section describes a second concept of interaction that perhaps brings us one step closer to the causal mechanisms by which treatments A and E bring about the outcome. This second concept of interaction is not based on counterfactual contrasts but rather on sufficient-component causes, and thus we refer to it as interaction within the sufficient-component-cause framework or, for brevity, *sufficient cause interaction*.

A sufficient cause interaction between A and E exists in the population if A and E occur together in a sufficient cause. For example, suppose individuals

Fine Point 5.2

From counterfactuals to sufficient-component causes, and vice versa. There is a correspondence between the counterfactual response types and the sufficient component causes. In the case of a dichotomous treatment and outcome, suppose an individual has none of the background factors U_0, U_1, U_2. She will have an "immune" response type because she lacks the components necessary to complete all of the sufficient causes, whether she is treated or not. The table below displays the mapping between response types and sufficient-component causes in the case of one treatment A.

Type	$Y^{a=0}$	$Y^{a=1}$	Component causes
Doomed	1	1	$U_0 = 1$ or $\{U_1 = 1$ and $U_2 = 1\}$
Helped	1	0	$U_0 = 0$ and $U_1 = 0$ and $U_2 = 1$
Hurt	0	1	$U_0 = 0$ and $U_1 = 1$ and $U_2 = 0$
Immune	0	0	$U_0 = 0$ and $U_1 = 0$ and $U_2 = 0$

A particular combination of component causes corresponds to one and only one counterfactual type. However, a particular response type may correspond to several combinations of component causes. For example, individuals of the "doomed" type may have any combination of component causes including $U_0 = 1$, no matter what the values of U_1 and U_2 are, or any combination including $\{U_1 = 1$ and $U_2 = 1\}$.

Sufficient-component causes can also be used to provide a mechanistic description of exchangeability $Y^a \perp\!\!\!\perp A$. For a dichotomous treatment and outcome, exchangeability means that the proportion of individuals who would have the outcome under treatment, and under no treatment, is the same in the treated $A = 1$ and the untreated $A = 0$. That is, $\Pr[Y^{a=1} = 1|A = 1] = \Pr[Y^{a=1} = 1|A = 0]$ and $\Pr[Y^{a=0} = 1|A = 1] = \Pr[Y^{a=0} = 1|A = 0]$.

Now the individuals who would develop the outcome if treated are the "doomed" and the "hurt", i.e., those with $U_0 = 1$ or $U_1 = 1$. The individuals who would get the outcome if untreated are the "doomed" and the "helped", that is, those with $U_0 = 1$ or $U_2 = 1$. Therefore there will be exchangeability if the proportions of "doomed" + "hurt" and of "doomed" + "helped" are equal in the treated and the untreated. That is, exchangeability for a dichotomous treatment and outcome can be expressed in terms of sufficient-component causes as $\Pr[U_0 = 1$ or $U_1 = 1|A = 1] = \Pr[U_0 = 1$ or $U_1 = 1|A = 0]$ and $\Pr[U_0 = 1$ or $U_2 = 1|A = 1] = \Pr[U_0 = 1$ or $U_2 = 1|A = 0]$.

For additional details see Greenland and Brumback (2002), Flanders (2006), and VanderWeele and Hernán (2006). Some of the above results were generalized to the case of two or more dichotomous treatments by VanderWeele and Robins (2008).

with background factors $U_5 = 1$ will develop the outcome when jointly receiving vitamins ($E = 1$) and heart transplant ($A = 1$), but not when receiving only one of the two treatments. Then a sufficient cause interaction between A and E exists if there exists an individual with $U_5 = 1$. It then follows that if there exists an individual with counterfactual responses $Y^{a=1,e=1} = 1$ and $Y^{a=0,e=1} = Y^{a=1,e=0} = 0$, a sufficient cause interaction between A and E is present.

Sufficient cause interactions can be synergistic or antagonistic. There is *synergism* between treatment A and treatment E when $A = 1$ and $E = 1$ are present in the same sufficient cause, and *antagonism* between treatment A and treatment E when $A = 1$ and $E = 0$ (or $A = 0$ and $E = 1$) are present in the same sufficient cause. Alternatively, one can think of antagonism between treatment A and treatment E as synergism between treatment A and no treatment E (or between no treatment A and treatment E).

Unlike the counterfactual definition of interaction, sufficient cause interaction makes explicit reference to the causal mechanisms involving the treatments A and E. One could then think that identifying the presence of sufficient cause interaction requires detailed knowledge about these causal mechanisms. It turns out that this is not always the case: sometimes we can conclude that

Fine Point 5.3

Biologic interaction. In epidemiologic discussions, sufficient-cause interaction is commonly referred to as biologic interaction (Rothman et al, 1980). This choice of terminology might seem to imply that, in biomedical applications, there exist biological mechanisms through which two treatments A and E act on each other in bringing about the outcome. However, this may not be necessarily the case as illustrated by the following example proposed by VanderWeele and Robins (2007a).

Suppose A and E are the two alleles of a gene that produces an essential protein. Individuals with a deleterious mutation in both alleles ($A = 1$ and $E = 1$) will lack the essential protein and die within a week after birth, whereas those with a mutation in none of the alleles (i.e., $A = 0$ and $E = 0$) or in only one of the alleles (i.e., $A = 0$ and $E = 1$, $A = 1$ and $E = 0$) will have normal levels of the protein and will survive. We would say that there is synergism between the alleles A and E because there exists a sufficient component cause of death that includes $A = 1$ and $E = 1$. That is, both alleles work together to produce the outcome. However, it might be argued that they do not physically act on each other and thus that they do not interact in any biological sense.

Rothman (1976) described the concepts of synergism and antagonism within the sufficient-component-cause framework.

sufficient cause interaction exists even if we lack any knowledge whatsoever about the sufficient causes and their components. Specifically, if the inequalities in Fine Point 5.1 hold, then there exists synergism between A and E. That is, one can empirically check that synergism is present without ever giving any thought to the causal mechanisms by which A and E work together to bring about the outcome. This result is not that surprising because of the correspondence between counterfactual response types and sufficient causes (see Fine Point 5.2), and because the above inequality is a sufficient but not a necessary condition, i.e., the inequality may not hold even if synergism exists.

5.6 Counterfactuals or sufficient-component causes?

A counterfactual framework of causation was already hinted at by Hume (1748).

The sufficient-component-cause framework was developed in philosophy by Mackie (1965). He introduced the concept of *INUS* condition for Y: an *I*nsufficient but *N*ecessary part of a condition which is itself *U*nnecessary but exclusively *S*ufficient for Y.

The sufficient-component-cause framework and the counterfactual (potential outcomes) framework address different questions. The sufficient-component-cause model considers sets of actions, events, or states of nature which together inevitably bring about the outcome under consideration. The model gives an account of the causes of a particular effect. It addresses the question, "Given a particular effect, what are the various events which might have been its cause?" The potential outcomes or counterfactual model focuses on one particular cause or intervention and gives an account of the various effects of that cause. In contrast to the sufficient-component-cause framework, the potential outcomes framework addresses the question, "What would have occurred if a particular factor were intervened upon and thus set to a different level than it in fact was?" Unlike the sufficient-component-cause framework, the counterfactual framework does not require a detailed knowledge of the mechanisms by which the factor affects the outcome.

The counterfactual approach addresses the question "what happens?" The sufficient-component-cause approach addresses the question "how does it happen?" For the contents of this book—conditions and methods to estimate the average causal effects of hypothetical interventions—the counterfactual framework is the natural one. The sufficient-component-cause framework is helpful to think about the causal mechanisms at work in bringing about a particular outcome. Sufficient-component causes have a rightful place in the teaching of

Fine Point 5.4

More on the attributable fraction. Fine Point 3.6 defined the excess fraction for treatment A as the proportion of cases attributable to treatment A in a particular population, and described an example in which the excess fraction for A was 75%. That is, 75% of the cases would not have occurred if everybody had received treatment $a = 0$ rather than their observed treatment A. Now consider a second treatment E. Suppose that the excess fraction for E is 50%. Does this mean that a joint intervention on A and E could prevent 125% ($75\% + 50\%$) of the cases? Of course not.

Clearly the excess fraction cannot exceed 100% for a single treatment (either A or E). Similarly, it should be clear that the excess fraction for any joint intervention on A and E cannot exceed 100%. That is, if we were allowed to intervene in any way we wish (by modifying A, E, or both) in a population, we could never prevent a fraction of disease greater than 100%. In other words, no more than 100% of the cases can be attributed to the lack of certain intervention, whether single or joint. But then why is the sum of excess fractions for two single treatments greater than 100%? The sufficient-component-cause framework helps answer this question.

As an example, suppose that Zeus had background factors $U_5 = 1$ (and none of the other background factors) and was treated with both $A = 1$ and $E = 1$. Zeus would not have been a case if either treatment A or treatment E had been withheld. Thus Zeus is counted as a case prevented by an intervention that sets $a = 0$, i.e., Zeus is part of the 75% of cases attributable to A. But Zeus is also counted as a case prevented by an intervention that sets $e = 0$, i.e., Zeus is part of the 50% of cases attributable to E. No wonder the sum of the excess fractions for A and E exceeds 100%: some individuals like Zeus are counted twice!

The sufficient-component-cause framework shows that it makes little sense to talk about the fraction of disease attributable to A and E separately when both may be components of the same sufficient cause. For example, the discussion about the fraction of disease attributable to either genes or environment is misleading. Consider the mental retardation caused by phenylketonuria, a condition that appears in genetically susceptible individuals who eat certain foods. The excess fraction for those foods is 100% because all cases can be prevented by removing the foods from their diet. The excess fraction for the genes is also 100% because all cases would be prevented if we could replace the susceptibility genes. Thus the causes of mental retardation can be seen as either 100% genetic or 100% environmental. See Rothman, Greenland, and Lash (2008) for further discussion.

causal inference because they help understand key concepts like the dependence of the magnitude of causal effects on the distribution of background factors (effect modifiers), and the relationship between effect modification, interaction, and synergism.

Though the sufficient-component-cause framework is useful from a pedagogic standpoint, its relevance to actual data analysis is yet to be determined. In its classical form, the sufficient-component-cause framework is deterministic, its conclusions depend on the coding of the outcome, and is by definition limited to dichotomous treatments and outcomes (or to variables that can be recoded as dichotomous variables). This limitation practically rules out the consideration of any continuous factors, and restricts the applicability of the framework to contexts with a small number of dichotomous factors. More recent extensions of the sufficient-component-cause framework to stochastic settings and to categorical and ordinal treatments might lead to an increased application of this approach to realistic data analysis. Finally, even allowing for these extensions of the sufficient-component-cause framework, we may rarely have the large amount of data needed to study the fine distinctions it makes.

VanderWeele (2010b) provided extensions to 3-level treatments. VanderWeele and Robins (2012) explored the relationship between stochastic counterfactuals and stochastic sufficient causes.

To estimate causal effects more generally, the counterfactual framework will likely continue to be the one most often employed. Some apparently alternative frameworks—causal diagrams, decision theory—are essentially equivalent to the counterfactual framework, as described in the next chapter.

Technical Point 5.3

Monotonicity of causal effects and sufficient causes. When treatment A and E have monotonic effects, then some sufficient causes are guaranteed not to exist. For example, suppose that cigarette smoking $(A = 1)$ never prevents heart disease, and that physical inactivity $(E = 1)$ never prevents heart disease. Then no sufficient causes including either $A = 0$ or $E = 0$ can be present. This is so because, if a sufficient cause including the component $A = 0$ existed, then some individuals (e.g., those with $U_2 = 1$) would develop the outcome if they were unexposed $(A = 0)$ or, equivalently, the outcome could be prevented in those individuals by treating them $(A = 1)$. The same rationale applies to $E = 0$. The sufficient component causes that cannot exist when the effects of A and E are monotonic are crossed out in Figure 5.3.

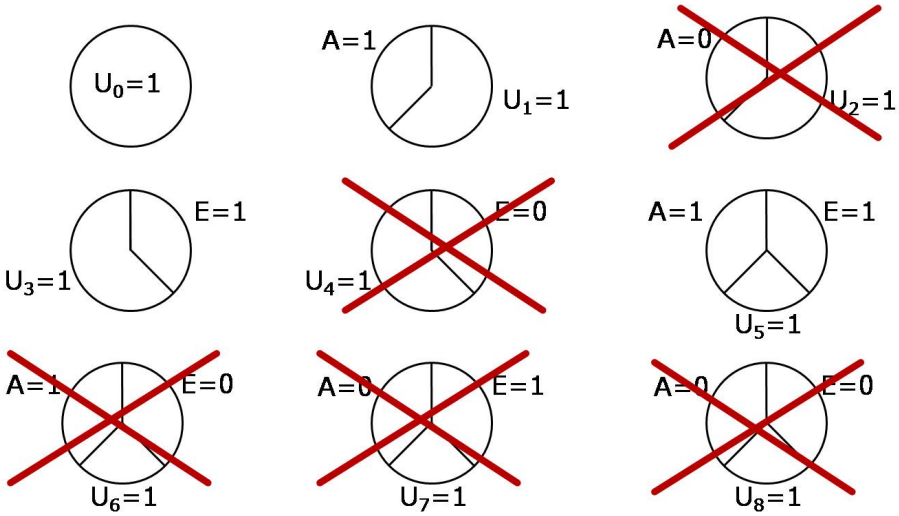


Figure 5.3