

# Identifiability and Exchangeability for Direct and Indirect Effects

James M. Robins<sup>1</sup> and Sander Greenland<sup>2</sup>

We consider the problem of separating the direct effects of an exposure from effects relayed through an intermediate variable (indirect effects). We show that adjustment for the intermediate variable, which is the most common method of estimating direct effects, can be biased. We also show that, even in a randomized crossover trial of exposure, direct and indirect effects cannot be separated without special assumptions; in other words, direct and indirect effects are not separately identifiable when only exposure is randomized. If the exposure and intermediate never interact to cause disease and if intermediate effects can be controlled, that is, blocked by a suitable intervention, then a trial randomizing both exposure and the intervention can separate direct from indirect effects. Nonetheless, the estimation must be carried

out using the G-computation algorithm. Conventional adjustment methods remain biased. When exposure and the intermediate interact to cause disease, direct and indirect effects will not be separable even in a trial in which both the exposure and the intervention blocking intermediate effects are randomly assigned. Nonetheless, in such a trial, one can still estimate the fraction of exposure-induced disease that could be prevented by control of the intermediate. Even in the absence of an intervention blocking the intermediate effect, the fraction of exposure-induced disease that could be prevented by control of the intermediate can be estimated with the G-computation algorithm if data are obtained on additional confounding variables. (Epidemiology 1992;3:143-155)

**Keywords:** causality, causal modeling, epidemiologic methods, risk.

Once it is established that changes in an exposure variable affect disease risk, questions often arise as to the relative importance of different possible pathways for the effect. For example, if one wishes to evaluate intervention on smoking in the prevention of cardiovascular disease, one confronts the difficulty of effecting changes in smoking habit. Thus, from a public health point of view, one might ask what fraction of smoking's effect could be eliminated by controlling (that is, eliminating) all hyperlipidemia through diet and/or medication. From a mechanistic point of view, one might ask what fraction of smoking's effect is "indirect," in the sense of being mediated through its effect on serum lipid levels, and what fraction of

smoking's effect on risk is "direct" relative to hyperlipidemia, in the sense of being mediated through pathways that do not involve serum lipid levels. We note that it is quite plausible that smoking and hyperlipidemia interact to produce clinical cardiovascular disease as, for example, if hyperlipidemia produced coronary artery stenosis, the stenotic artery was blocked by a thrombus caused by smoking-induced platelet aggregation, and the thrombosis resulted in a myocardial infarction.

We will show that when smoking and the intermediate hyperlipidemia do not interact to cause disease, the fraction of smoking-induced disease that could be prevented by controlling hyperlipidemia will equal the fraction of disease attributable to the indirect effect of smoking. In the presence of interaction, these fractions will differ.

In Section 1, we show that a common approach to estimating direct effects is often biased, in that it can yield confounded estimates in both observational epidemiologic studies and conventional randomized trials. In Section 2, we show that, even in a randomized crossover trial of exposure, direct and indirect effects cannot be separated without special assumptions. In Sections 3 and 4, we show that if the exposure and the intermediate do not interact and if the intermediate effects can be controlled, that is, blocked by a suitable

From the <sup>1</sup>Occupational Health Program and Department of Biostatistics, Harvard School of Public Health, Boston, MA, and <sup>2</sup>Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA.

Address correspondence to: James M. Robins, Occupational Health Program and Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115.

Supported by National Institute of Environmental Health Sciences Grants 5-K04-ES00180, P30-ES00002, and R01-ES034-05 [J. M. R.], and by National Institutes of Health Contract N01-AI-72631 [S. G.].

© 1992 Epidemiology Resources Inc.

intervention, then a trial randomizing both exposure and the intervention can separate direct from indirect effects. The estimation must be carried out using the G-computation algorithm. Conventional adjustment methods remain biased. Furthermore, we show that even in the absence of an intervention blocking the intermediate effect, direct and indirect effects can be separated with the G-computation algorithm, provided that data are obtained on additional confounding variables.

In Section 5, we show that when exposure and the intermediate do interact to cause disease, direct and indirect effects will not be separable even in the trial in which both the exposure and the intervention blocking intermediate effects are randomly assigned. Nonetheless, in such a trial, one can still use the G-computation algorithm to estimate the fraction of exposure-induced disease that could be prevented by control of the intermediate.

Our results have implications for the design, analysis, and interpretation of observational studies. Specifically, when data are obtained on important confounding variables, one can use the G-computation algorithm to estimate both the total (that is, net) exposure effect and the fraction of that effect that could be eliminated by control of the intermediate. In the absence of interaction between exposure and intermediate, this same approach will allow separation of direct and indirect effects. In contrast, if exposure and the intermediate interact to cause disease, direct and indirect exposure effects will not be separable even if data on confounders are obtained.

We wish to emphasize the difference between the subject matter addressed in this paper and that addressed in Ref 1, in which Robins considered the estimation of the total (that is, net) effect of a time-varying exposure in the presence of covariates that were simultaneously confounders for the total exposure effect and intermediate variables. In contrast, in this paper, we consider the estimation of the direct and indirect effects of a time-independent exposure.

### 1. An Elementary Causal Model for Direct and Indirect Effects

In this section, we will use an extension developed by Robins<sup>1-7</sup> of a basic causal model we have used elsewhere<sup>8-14</sup> to illustrate the potential bias in estimating direct effects by adjusting for the potential intermediate. Indeed, we will show that direct and indirect effects may not be separately identified even in a crossover randomized trial of exposure. The basic model has a long history in philosophy and statistics

and is sometimes known as a "counterfactual model"<sup>15</sup> or Rubin's model.<sup>16</sup> Elements of it can be found in works by Fisher and Neyman from the 1920s<sup>17</sup>; see Refs 16-18 for further details. The model can be extended to incorporate continuous covariates and outcomes (such as survival times), time-dependent covariates, and random outcomes; Refs 1-7 and 11-14 provide examples. We here limit our examples to dichotomous variables and deterministic outcomes; this limitation introduces a certain degree of artificiality into the examples, but it makes the computations transparent.

Consider an exposure or treatment variable  $X$ , disease variable  $D$ , and covariate  $Z$ , all coded 1 = "occurs," 0 = "does not occur"; level 1 of  $X$  represents the study exposure, level 1 of  $D$  represents the study disease, and level 1 of  $Z$  represents the potentially intermediate cofactor. For concreteness, suppose we have a large cohort of male smokers who have agreed to quit smoking if they are randomized to a cessation program, and who will continue smoking otherwise. Let  $X = 0$  for those randomized to quit smoking and  $X = 1$  for those randomized to continue; for simplicity, we will assume that all subjects comply with their assigned treatment, although this assumption would not be essential for the general theory if one were interested in assigned treatment as the study exposure, rather than actual treatment.<sup>5</sup> Let  $D = 1$  represent cardiovascular disease, 0 no cardiovascular disease; and let  $Z = 1$  represent hyperlipidemia, 0 normal serum lipids. We suppose that, at baseline, all subjects have normal lipids and no cardiovascular disease. To avoid temporal ambiguity, we shall suppose that, for all subjects, a single serum lipid measurement is made at a time  $t_1$  after randomization, no subject develops cardiovascular disease before  $t_1$ , and the outcome of interest is the development of cardiovascular disease by the end of follow-up at time  $t_2$ .

Fundamental to our causal model is the notion of a *counterfactual* conditional statement.<sup>15-17</sup> Each subject in the cohort is observed under one set of circumstances, but we also consider what would have happened to the subject under so-called counterfactual circumstances—circumstances that, contrary to fact, did not occur. For example, to define effects of smoking, we wish to consider (a) what actually happened to smoking subjects, and (b) what would have happened to smoking subjects if, contrary to fact, they had quit smoking.

Specifically, for each subject, we wish to consider whether hyperlipidemia would occur if he smokes and whether hyperlipidemia would occur if he quits; we

also wish to consider whether cardiovascular disease would occur under each possible combination of smoking and serum lipid status. To illustrate, consider a subject who continued smoking ( $X = 1$ ), then developed hyperlipidemia ( $Z = 1$ ), and then developed cardiovascular disease ( $D = 1$ ). The variable that indicates what the subject's serum lipid status would have been if, contrary to fact, he had quit smoking is a counterfactual variable; also counterfactual are this subject's disease status if (a) he had continued to smoke but had not developed hyperlipidemia, (b) he had quit smoking but had developed hyperlipidemia, and (c) he had quit smoking and had not developed hyperlipidemia. Despite their counterfactual nature, these three additional outcome variables are useful for deducing the effects of interventions.

To aid in these deductions, we will cross-classify each subject in the cohort into a  $2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^6$  table, according to whether or not the following statements are true or false for the subject; each statement is a hypothesis about the response of the subject to various (possibly counterfactual) combinations of smoking and serum lipid status:

- 1) Hyperlipidemia would occur if smoking continues (denoted  $Z = 1 \mid X = 1$ ).
- 2) Hyperlipidemia would occur if smoking ceases ( $Z = 1 \mid X = 0$ ).
- 3) Cardiovascular disease would occur if smoking continues and hyperlipidemia occurs ( $D = 1 \mid X = Z = 1$ ).
- 4) Cardiovascular disease would occur if smoking continues but lipids remain normal ( $D = 1 \mid X = 1, Z = 0$ ).
- 5) Cardiovascular disease would occur if smoking ceases but hyperlipidemia occurs ( $D = 1 \mid X = 0, Z = 1$ ).
- 6) Cardiovascular disease would occur if smoking ceases and lipids remain normal ( $D = 1 \mid X = Z = 0$ ).

Effects of continuing to smoke and of hyperlipidemia are defined according to combinations of these statements. For example, for a given continuing smoker, we say that hyperlipidemia is an effect of smoking (or smoking caused hyperlipidemia) if Statement 1 is true but Statement 2 is false (that is, if hyperlipidemia would occur if and only if the subject smoked). Symmetrically, we would say that smoking prevents hyperlipidemia in a smoker if Statement 2 is true but Statement 1 is false (that is, if the subject gets hyperlipidemia if and only if he stops smoking).

In theory, a subject could be classified as one of  $2^6 = 64$  possible types of subject according to which combination of the above six statements is true for the subject. For simplicity, however, we will assume throughout that smoking cannot prevent hyperlipidemia, and neither smoking nor hyperlipidemia can pre-

vent cardiovascular disease. This assumption means that (a) there are no subjects who develop hyperlipidemia if and only if they quit smoking; (b) there are no subjects who develop cardiovascular disease if and only if they quit smoking; and (c) there are no subjects who develop cardiovascular disease if and only if they do not develop hyperlipidemia. In other words, we assume that certain cells of the  $2^6$  table are empty: (a) cells for which Statement 1 is false but Statement 2 is true are empty; (b) cells for which Statement 6 is true but Statement 4 is false are empty, as are cells for which Statement 5 is true but Statement 3 is false; and (c) cells for which Statement 6 is true but Statement 5 is false are empty, as are cells for which Statement 4 is true but Statement 3 is false. These assumptions reduce the number of possible types to 18.

Until Section 5, we will also assume that smoking and hyperlipidemia never compete or interact to produce disease. This assumption means that (a) there are no subjects who develop cardiovascular disease if and only if they both continue smoking and develop hyperlipidemia; and (b) there are no subjects who develop cardiovascular disease if and only if they either continue smoking or develop hyperlipidemia. In other words, we assume that (a) cells for which Statement 3 is true but Statements 4–6 are false are empty; and (b) cells for which Statement 6 is false but Statements 3–5 are true are empty. These assumptions further reduce the number of possible causal types to 12. These 12 types are described in Table 1.

Table 2 gives a formal characterization of these 12 types in terms of their outcomes with respect to the occurrence of hyperlipidemia given smoking and non-smoking, and occurrence of cardiovascular disease under the four possible smoking-serum lipid combinations. A "1" in a column of Table 2 corresponds to the presence of a characteristic, whereas a "0" corresponds to its absence. As we will discuss below, the numbers in parentheses cannot be determined, even in a crossover trial, unless one intervenes to alter the serum lipid status of subjects.

From the descriptions in Tables 1 and 2, note that

- a) For Types 3, 5, and 7, smoking would be a direct cause of cardiovascular disease.
- b) For Types 2, 4, 5, and 9, smoking would be a cause of hyperlipidemia.
- c) For Type 4, smoking would be an indirect cause of cardiovascular disease.
- d) For Types 1, 6, 8, 10, and 11, smoking would have no effect on either cardiovascular disease or hyperlipidemia.

If we denote the proportion of the cohort of causal type  $j$  ( $j = 0, \dots, 11$ ) by  $p_j$ , it follows that smoking can

TABLE 1. Description of 12 Causal Types of Subjects

| Type | Description                                                                                                                                                                                                                        |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 0    | Hyperlipidemia occurs, unaffected by smoking, and cardiovascular disease occurs, unaffected by smoking or hyperlipidemia (Statements 1-6 true)                                                                                     |
| 1    | Hyperlipidemia occurs, unaffected by smoking, and causes cardiovascular disease (Statements 1-3, 5 true; 4, 6 false)                                                                                                               |
| 2    | Smoking would cause hyperlipidemia; cardiovascular disease occurs, unaffected by smoking or hyperlipidemia (Statements 1, 3-6 true; 2 false)                                                                                       |
| 3    | Smoking would cause cardiovascular disease directly; hyperlipidemia occurs, unaffected by smoking, but has no effect (Statements 1-4 true; 5, 6 false)                                                                             |
| 4    | Smoking would cause hyperlipidemia, which in turn would cause cardiovascular disease (Statements 1, 3, 5 true; 2, 4, 6 false)                                                                                                      |
| 5    | Smoking would cause hyperlipidemia, and smoking would cause cardiovascular disease directly; hyperlipidemia has no effect (Statements 1, 3, 4 true; 2, 5, 6 false)                                                                 |
| 6    | Cardiovascular disease occurs, unaffected by smoking; hyperlipidemia does not occur (Statements 3-6 true; 1, 2 false)                                                                                                              |
| 7    | Smoking would cause cardiovascular disease directly; hyperlipidemia does not occur and would have no effect if it did (Statements 3, 4 true; 1, 2, 5, 6 false)                                                                     |
| 8    | Hyperlipidemia occurs, unaffected by smoking; cardiovascular disease does not occur, and neither hyperlipidemia nor smoking would cause disease (Statements 1, 2 true; 3-6 false)                                                  |
| 9    | Smoking would cause hyperlipidemia; cardiovascular disease does not occur, and neither smoking nor hyperlipidemia would cause cardiovascular disease (Statement 1 true; 2-6 false)                                                 |
| 10   | Smoking has no effect on hyperlipidemia or disease; neither hyperlipidemia nor cardiovascular disease occurs, although hyperlipidemia, if it occurred, would cause cardiovascular disease (Statements 3, 5 true; 1, 2, 4, 6 false) |
| 11   | Neither hyperlipidemia nor cardiovascular disease occurs, and neither hyperlipidemia nor smoking would cause cardiovascular disease (Statements 1-6 false)                                                                         |

have no direct effects on cardiovascular disease if  $p_3 = p_5 = p_7 = 0$ , and can have no indirect effects if  $p_4 = 0$ .

From Table 2, we can compute the expected proportions of each type of subject in each of the four outcome categories ( $Z = D = 1$ ;  $Z = 1, D = 0$ ;  $Z = 0, D = 1$ ; and  $Z = D = 0$ ). Table 3 gives these expected

proportions. These proportions are not observable, however; instead (subject to random variation), we observe only the smoking-specific incidence of each combination of  $Z$  and  $D$ , which are the column totals in Table 3. To understand how Table 3 was constructed, consider Type 5 subjects. Randomization

TABLE 2. Occurrence of Hyperlipidemia and Cardiovascular Disease Events under Different Conditions for the Causal Types in Table 1

| Type | Statement Number (See Text) |                     |                                    |                   |                     |           |
|------|-----------------------------|---------------------|------------------------------------|-------------------|---------------------|-----------|
|      | 1                           | 2                   | 3                                  | 4                 | 5                   | 6         |
|      | Serum Lipid Status If:      |                     | Cardiovascular Disease Status* If: |                   |                     |           |
|      | Smoker ( $X = 1$ )          | Quitter ( $X = 0$ ) | Smoker ( $X = 1$ )                 |                   | Quitter ( $X = 0$ ) |           |
|      |                             | Hyper ( $Z = 1$ )   | Normal ( $Z = 0$ )                 | Hyper ( $Z = 1$ ) | Normal ( $Z = 0$ )  |           |
| 0    | $Z = 1$                     | $Z = 1$             | $D = 1$                            | $D = (1)$         | $D = 1$             | $D = (1)$ |
| 1    | 1                           | 1                   | 1                                  | (0)               | 1                   | (0)       |
| 2    | 1                           | 0                   | 1                                  | (1)               | (1)                 | 1         |
| 3    | 1                           | 1                   | 1                                  | (1)               | 0                   | (0)       |
| 4    | 1                           | 0                   | 1                                  | (0)               | (1)                 | 0         |
| 5    | 1                           | 0                   | 1                                  | (1)               | (0)                 | 0         |
| 6    | 0                           | 0                   | (1)                                | 1                 | (1)                 | 1         |
| 7    | 0                           | 0                   | (1)                                | 1                 | (0)                 | 0         |
| 8    | 1                           | 1                   | 0                                  | (0)               | 0                   | (0)       |
| 9    | 1                           | 0                   | 0                                  | (0)               | (0)                 | 0         |
| 10   | 0                           | 0                   | (1)                                | 0                 | (1)                 | 0         |
| 11   | 0                           | 0                   | (0)                                | 0                 | (0)                 | 0         |

In body of table, 1 = statement true (event occurs), and 0 = statement false (event does not occur). Hyper = hyperlipidemia ( $Z = 1$ ), Normal = normal serum lipids ( $Z = 0$ ).

\* Outcomes in parentheses are counterfactual (and so remain unobserved) if there is no experimental manipulation of serum lipid status, even in a double-blind crossover trial of smoking.

TABLE 3. Expected Proportions of Subjects in Each Smoking Cessation Group for Each Combination of Type, Hyperlipidemia Status, and Cardiovascular Disease Status, under Randomization of Smoking Cessation ( $X = 0$  for Quitters, 1 for Continuing Smokers)

| Type      | Smokers                           |               |                                   |               | Quitters                          |               |                                   |               |
|-----------|-----------------------------------|---------------|-----------------------------------|---------------|-----------------------------------|---------------|-----------------------------------|---------------|
|           | Hyper ( $Z = 1$ )                 |               | Normal ( $Z = 0$ )                |               | Hyper ( $Z = 1$ )                 |               | Normal ( $Z = 0$ )                |               |
|           | $D = 1$                           | $D = 0$       | $D = 1$                           | $D = 0$       | $D = 1$                           | $D = 0$       | $D = 1$                           | $D = 0$       |
| 0         | $p_0$                             |               |                                   |               | $p_0$                             |               |                                   |               |
| 1         | $p_1$                             |               |                                   |               | $p_1$                             |               |                                   |               |
| 2         | $p_2$                             |               |                                   |               |                                   |               | $p_2$                             |               |
| 3         | $p_3$                             |               |                                   |               |                                   | $p_3$         |                                   |               |
| 4         | $p_4$                             |               |                                   |               |                                   |               |                                   | $p_4$         |
| 5         | $p_5$                             |               |                                   |               |                                   |               |                                   | $p_5$         |
| 6         |                                   |               | $p_6$                             |               |                                   |               | $p_6$                             |               |
| 7         |                                   |               | $p_7$                             |               |                                   |               |                                   | $p_7$         |
| 8         |                                   | $p_8$         |                                   |               |                                   | $p_8$         |                                   |               |
| 9         |                                   | $p_9$         |                                   |               |                                   |               |                                   | $p_9$         |
| 10        |                                   |               |                                   | $p_{10}$      |                                   |               |                                   | $p_{10}$      |
| 11        |                                   |               |                                   | $p_{11}$      |                                   |               |                                   | $p_{11}$      |
| Subtotals | $\pi_{11}$                        | $\sigma_{11}$ | $\pi_{10}$                        | $\sigma_{10}$ | $\pi_{01}$                        | $\sigma_{01}$ | $\pi_{00}$                        | $\sigma_{00}$ |
| Totals    | $I_{11} = \pi_{11} + \sigma_{11}$ |               | $I_{10} = \pi_{10} + \sigma_{10}$ |               | $I_{01} = \pi_{01} + \sigma_{01}$ |               | $I_{00} = \pi_{00} + \sigma_{00}$ |               |

Note:  $I_{11} + I_{10} = 1$ ;  $I_{01} + I_{00} = 1$ .  
 Note that only the column totals ( $\pi_{jk}$ ,  $\sigma_{jk}$ , and  $I_{jk}$ ) are observable.  $D = 1$  if cardiovascular disease occurs, 0 otherwise;  $I_{11}$  and  $I_{01}$  are the expected incidences of hyperlipidemia among smokers and quitters. Blank entries represent impossible combinations in the examples.

guarantees that the expected proportion of Type 5 subjects among smokers would equal that among non-smokers: both would equal  $p_5$ . Type 5 subjects who smoke would develop hyperlipidemia ( $Z = 1$ ) and cardiovascular disease ( $D = 1$ ). In contrast, unexposed Type 5 subjects will be normolipidemic ( $Z = 0$ ) and thus will not develop cardiovascular disease ( $D = 0$ ). Table 4 rearranges these totals in the more familiar  $2 \times 2 \times 2$  table format, and Table 5 displays the incidence proportions (average risks) in terms of the proportions of causal types (the  $p_i$ ). We note four points about these incidence proportions:

- 1) The total excess disease incidence (risk difference of cardiovascular disease) due to smoking (which we call the net, total, or overall effect of exposure) is  $p_3 + p_4 + p_5 +$

- $p_7$ ; because exposure was randomized, this sum equals the crude risk difference  $R_{1c} - R_{0c}$ .
- 2) The excess cardiovascular disease incidence due to direct smoking effects is  $p_3 + p_5 + p_7$ .
- 3) The excess cardiovascular disease incidence due to indirect smoking effects (through hyperlipidemia) is  $p_4$ .
- 4) The sum of the quantities in Points 2 and 3 is the quantity in Point 1: if only the 12 types in Tables 1 and 2 are present, the total smoking effect equals the sum of the direct and indirect effects.

As we discuss later, Point 4 is not true in general: if there are subjects for which smoking causes hyperlipidemia and then interacts with hyperlipidemia to cause cardiovascular disease, the sum of the direct and indirect effects can exceed the total excess incidence.

The following example shows that, without special assumptions, neither the excess cardiovascular disease incidence due to direct effects nor the excess cardiovascular disease incidence due to indirect effects can be validly estimated or even tested by adjusting the smoking effect for serum lipid status.

EXAMPLE 1

Suppose there are no direct effects, so that  $p_3 = p_5 = p_7 = 0$ , and that  $p_0 = 0.112$ ,  $p_1 = 0.012$ ,  $p_2 = 0.028$ ,  $p_4 = 0.048$ ,  $p_6 = 0.140$ ,  $p_8 = 0.252$ ,  $p_9 = 0.120$ ,  $p_{10} = 0.060$ , and  $p_{11} = 0.228$ . Then, computing directly from Panel II of Table 5, the expected proportions getting cardiovascular disease (incidence proportions) will appear as follows:

TABLE 4. Expected Numbers for Data Observable in a Randomized Trial of Smoking Cessation When  $N_1$  Are Randomized to No Treatment ( $X = 1$ ) and  $N_0$  Are Randomized to Quitting ( $X = 0$ )

|           | Hyperlipidemia                     |                                    | Normal                             |                                    |
|-----------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
|           | Smoker                             | Quitter                            | Smoker                             | Quitter                            |
| $D = 1$   | $\pi_{11}N_1$                      | $\pi_{01}N_0$                      | $\pi_{10}N_1$                      | $\pi_{00}N_0$                      |
| $D = 0$   | $\sigma_{11}N_1$                   | $\sigma_{01}N_0$                   | $\sigma_{10}N_1$                   | $\sigma_{00}N_0$                   |
| Totals    | $I_{11}N_1$                        | $I_{01}N_0$                        | $I_{10}N_1$                        | $I_{00}N_0$                        |
| Incidence | $R_{11} = \frac{\pi_{11}}{I_{11}}$ | $R_{01} = \frac{\pi_{01}}{I_{01}}$ | $R_{10} = \frac{\pi_{10}}{I_{10}}$ | $R_{00} = \frac{\pi_{00}}{I_{00}}$ |

**TABLE 5. Expected Incidence of Hyperlipidemia ( $Z = 1$ ) and Cardiovascular Disease ( $D = 1$ ) under Randomization of All Cohort Members to Smoking Cessation ( $X = 0$  for Quitters, 1 for Continuing Smokers)**

|                                                                          |                                                                                                                              |
|--------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| I. Expected proportions experiencing hyperlipidemia when                 |                                                                                                                              |
| $X = 1$                                                                  | $I_{11} = \pi_{11} + \sigma_{11} = p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_8 + p_9$                                            |
| $X = 0$                                                                  | $I_{01} = \pi_{01} + \sigma_{01} = p_0 + p_1 + p_3 + p_8$                                                                    |
| II. Expected proportions experiencing cardiovascular disease when        |                                                                                                                              |
| $Z = 1$ and $X = 1$                                                      | $R_{11} = \frac{\pi_{11}}{I_{11}} = \frac{p_0 + p_1 + p_2 + p_3 + p_4 + p_5}{p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_8 + p_9}$ |
| $Z = 1$ and $X = 0$                                                      | $R_{01} = \frac{\pi_{01}}{I_{01}} = \frac{(p_0 + p_1)}{(p_0 + p_1 + p_3 + p_8)}$                                             |
| $Z = 0$ and $X = 1$                                                      | $R_{10} = \frac{\pi_{10}}{I_{10}} = \frac{(p_6 + p_7)}{(p_6 + p_7 + p_{10} + p_{11})}$                                       |
| $Z = 0$ and $X = 0$                                                      | $R_{00} = \frac{\pi_{00}}{I_{00}} = \frac{(p_2 + p_6)}{(p_2 + p_4 + p_5 + p_6 + p_7 + p_9 + p_{10} + p_{11})}$               |
| III. Crude expected proportions experiencing cardiovascular disease when |                                                                                                                              |
| $X = 1$                                                                  | $R_{1c} = \pi_{11} + \pi_{10} = I_{11}R_{11} + I_{10}R_{10} = p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7$                 |
| $X = 0$                                                                  | $R_{0c} = \pi_{01} + \pi_{00} = I_{01}R_{01} + I_{00}R_{00} = p_0 + p_1 + p_2 + p_6$                                         |

$p_i$  is the proportion of cohort members of causal type  $i$  ( $i = 0, \dots, 11$ ), described in Table 1;  $\pi_{jk}$ ,  $\sigma_{jk}$ , and  $I_{jk}$  are the observable proportions defined in Tables 3 and 4.

| Hyperlipidemia  | Continued smoking | Expected cardiovascular disease incidence |
|-----------------|-------------------|-------------------------------------------|
| Yes ( $Z = 1$ ) | Yes ( $X = 1$ )   | $R_{11} = 0.200/0.572 = 0.350$            |
|                 | No ( $X = 0$ )    | $R_{01} = 0.124/0.376 = 0.330$            |
| No ( $Z = 0$ )  | Yes ( $X = 1$ )   | $R_{10} = 0.140/0.428 = 0.327$            |
|                 | No ( $X = 0$ )    | $R_{00} = 0.168/0.624 = 0.269$            |

Within both levels of serum lipids, smoking is associated with an increased risk of cardiovascular disease, so that any lipid-adjusted estimate of the smoking-cardiovascular disease association (such as the Mantel-Haenszel odds ratio or a logistic regression coefficient) will in expectation show a positive exposure effect, even though there are no direct effects of smoking on cardiovascular disease in our example. Likewise, an adjusted test of the null hypothesis of no direct effects (such as the Mantel-Haenszel test) will be invalid, in that the probability of rejection will approach certainty as the sample size increases. Furthermore, the expected adjusted estimate of effect is close to the expected crude estimate, so that one might be tempted to infer (incorrectly) that most or all of smoking's effect is direct. For example, using the total cohort as standard, the expected standardized risk difference would be  $0.040$ , close to the expected crude risk difference  $R_{1c} - R_{0c} = 0.340 - 0.292 = 0.048$ .

Not only is the conventional method of estimating direct effects (that is, adjusting for the intermediate covariate) biased in the above example, but without further constraints it is impossible to obtain a valid estimate of either the direct or the indirect effects. In contrast, because smoking is randomized, the net effect

of smoking  $p_3 + p_4 + p_5 + p_7$  is unbiasedly estimated by the crude risk difference.<sup>17,18</sup>

### 2. Nonidentifiability in a Crossover Trial

Under the model given above or under more complex models, we could still not separate direct and indirect effects even if we could conduct a perfect crossover trial in which there were no carryover effects (that is, in which the results do not depend on the time order of the exposed and nonexposed periods of the trial). In such a trial, we would observe each subject's responses to exposure and to nonexposure, yet we still could not identify the causal type of each subject. In particular, it can be seen from Table 2 that Type 4 and Type 5 subjects would display identical responses to smoking ( $X = 1$ ) and to quitting ( $X = 0$ ): both types would develop hyperlipidemia and cardiovascular disease if and only if they continued to smoke, so these types would be indistinguishable. Unfortunately, Type 4 subjects represent indirect effects ( $p_4$ ), whereas Type 5 subjects represent part of the direct effects ( $p_3 + p_5 + p_7$ ). Thus, even in this ideal setting, direct and indirect effects could not be separated. Assuming  $p_5 = 0$  (that is, no Type 5 individuals) would render these effects separable in the crossover trial with no carryover effects, but such an assumption would rarely be justified based on available biological knowledge.

### 3. Randomization of Cofactor Intervention

A crossover trial is rarely possible, and, as just shown, even such an ideal design will not allow separation of

direct and indirect effects without additional assumptions. We now examine some conditions sufficient for separation of effects when the intermediate variable (as well as the exposure) can be manipulated.

It is sometimes feasible to control, that is, block, the effect of the intermediate cofactor on cardiovascular disease via some intervention. For example, hyperlipidemia may be treated with drugs or diet. Randomization of such interventions within exposure levels can allow separation of direct and indirect exposure effects if only the types in Table 1 are present and the intervention is completely effective in preventing effects of the intermediate cofactor.

To illustrate this point, assume that the causal model in Tables 1 and 2 holds, and we have a feasible serum lipid intervention with the following properties:

- 1) The intervention prevents all hyperlipidemia effects but has no effect on cardiovascular disease risk other than through its effect on restoring normal lipid levels; and
- 2) Aside from the intervention's effect, intervention compliance is not predictive of cardiovascular disease risk.

These two assumptions may be weakened to be conditional on any baseline prognostic factors that are controlled in the analysis. In a trial in which (a) smoking cessation was randomized, and (b) the cofactor intervention is randomly allocated to subjects in whom hyperlipidemia occurs, the expected cardiovascular disease incidences among subjects who do not receive the intervention would appear as in Table 5. Among subjects who do receive the intervention, however, and thus experience no hyperlipidemia effects, the expected incidence among the exposed,  $R_{11B}$ , will be

$$\frac{(p_0 + p_2 + p_3 + p_5)}{(p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_8 + p_9)} \quad (1)$$

Eq 1 is obtained by dropping  $p_1$  and  $p_4$  from the numerator of  $R_{11}$ , because Type 1 and Type 4 subjects will not get disease if hyperlipidemia effects are prevented. Similarly, the expected incidence among quitters who receive the intervention,  $R_{01B}$ , will be

$$p_0 / (p_0 + p_1 + p_3 + p_8); \quad (2)$$

this expression is obtained by dropping  $p_1$  from the numerator of  $R_{01}$ . The two incidences  $R_{11B}$  and  $R_{01B}$ , combined with those in Table 5, allow one to estimate the indirect effect  $p_4$ , since the latter equals the estimable quantity

$$I_{11}(R_{11} - R_{11B}) - I_{01}(R_{01} - R_{01B}); \quad (3)$$

this expression is obtained by noting that  $I_{11}R_{11} = p_0 + p_1 + p_2 + p_3 + p_4 + p_5$ ,  $I_{11}R_{11B} = p_0 + p_2 + p_3 +$

$p_5$ ,  $I_{01}R_{01} = p_0 + p_1$ , and  $I_{01}R_{01B} = p_0$ . The direct effect of smoking,  $p_3 + p_5 + p_7$ , can be estimated as the net effect minus the indirect effect,

$$R_{1c} - R_{0c} - [I_{11}(R_{11} - R_{11B}) - I_{01}(R_{01} - R_{01B})]. \quad (4)$$

EXAMPLE 2

Consider again Example 1, in which there were no direct effects of smoking. We have  $R_{11B} = 0.140/0.572 = 0.245$  and  $R_{01B} = 0.112/0.376 = 0.298$ . Thus, from Eq 3, the estimate of the indirect effect  $p_4$  is

$$0.572(0.350 - 0.245) - 0.376(0.330 - 0.298) = 0.048,$$

and, from Eq 4, the estimate of the direct effect is  $0.048 - 0.048 = 0.000$ ; both estimates are correct. Nevertheless, we have seen in Section 1 that simply stratifying the observed data on hyperlipidemia status would yield a biased estimate. Eqs 3 and 4 are examples of the G-computation algorithm described in Refs 1-6.

4. Exchangeability Assumptions for Separation of Effects

In the absence of randomization of the intermediate cofactor (hyperlipidemia), it is natural to estimate  $R_{11B}$  and  $R_{01B}$  from the estimates of  $R_{10}$  and  $R_{00}$ . This substitution would yield an unbiased estimate if  $R_{11B} = R_{10}$  and  $R_{01B} = R_{00}$ . The latter conditions may be restated as the following "partial exchangeability" assumptions:

- E1) The expected cardiovascular disease incidence among the smoking normolipidemics equals the incidence that the smoking hyperlipidemics would have had if their hyperlipidemia had been prevented;
- E2) The expected cardiovascular disease incidence among the quitting normolipidemics equals the incidence that the quitting hyperlipidemics would have had if their hyperlipidemia had been prevented.

EXAMPLE 3

In Example 2,  $R_{11B} = 0.245 \neq 0.327 = R_{10}$  and  $R_{01B} = 0.298 \neq 0.269 = R_{00}$ , so that E1 and E2 are violated: smoking normolipidemics have a higher risk than smoking hyperlipidemics would have if their hyperlipidemia was prevented, and quitting normolipidemics have lower risk than quitting hyperlipidemics would have if their hyperlipidemia was prevented. Substitution of  $R_{10}$  and  $R_{00}$  for  $R_{11B}$  and  $R_{01B}$  in Eq 4 thus yields the biased value of  $R_{10} - R_{00} = 0.327 - 0.269 = 0.061$  for the direct effect. If E1 and E2 had been true, smoking were randomly assigned, and smoking and

hyperlipidemia do not interact to cause disease, then the stratum-specific risk difference in both the hyperlipidemics and normolipidemics would have been equal to the direct effect of exposure, that is,  $R_{11} - R_{01} = R_{10} - R_{00} = p_3 + p_5 + p_7$ .

Without randomization, one must fall back on certain exchangeability assumptions to identify the effects of interest.<sup>6</sup> To identify the net (total) effect of smoking in the single  $2 \times 2$  table that ignores hyperlipidemia, we must assume that the exposed and unexposed groups are exchangeable (comparable), that is, that the cardiovascular disease incidence among quitters equals the incidence that the smokers would have had if they had quit. Note, however, that to identify direct and indirect effects, one needs E1 and E2 in addition to the assumption of exchangeability of the exposed and nonexposed. Thus, one can have confounding for the direct effects but not the net effects. (It is also theoretically possible to have confounding of the net effects but not the direct effects.<sup>1-3</sup>)

EXPLANATORY COVARIATES (CONFOUNDERS)

Let us return now to our initial study design, in which only the exposure (continued smoking) was randomized, and in which stratification on the intermediate (hyperlipidemia) produced a biased (confounded) estimate of direct effect because the exchangeability assumptions E1 and E2 were violated. Ordinarily, one attempts to explain confounding by attributing nonexchangeability to the effects of additional covariates that predict risk (given exposure status) and are imbalanced across comparison groups. We will show that it is possible to explain the phenomenon we describe in these terms; nevertheless, one should note three properties of such explanations:

1. Such explanations are not unique, in that there are potentially an infinite number of ways in which uncontrolled covariates may be related to the observed study variables (here, smoking, hyperlipidemia, and cardiovascular disease) and produce exactly the same bias and the same observed distribution of the observed study variables. This property is shared by confounding in the estimation of net effects<sup>8</sup>; for example, a moderate degree of confounding might be produced by exposure being strongly associated with a moderate risk factor or weakly associated with a strong risk factor.
2. For estimating direct and indirect effects, the confounders may be postexposure covariates and thus may themselves be affected by exposure. In particular, the confounders may themselves be biological intermediates that precede the study intermediate (here, hyperlipidemia) in the causal pathway from exposure to disease; or they may be selection effects of exposure, such as leaving work.<sup>1-4</sup>
3. If the confounders are identified and measured but are

themselves affected by exposure, then simple stratification on the confounders will generally not suffice to control confounding; instead, one will have to resort to the G-computation algorithm,<sup>1-6</sup> special applications of which are given by Eq 3 above and Eq 5 below.

Properties 2 and 3 do not arise in the simple case of estimating the net effect of a point exposure, but they do arise in estimating the effects of a sustained exposure.<sup>1-7</sup>

The general form of the G-computation algorithm requires a more complex structure than we have developed here, but an illustration of Properties 2 and 3 above is given in the following extension of Example 1 to include a three-level covariate  $C = 0, 1, \text{ or } 2$ , which smoking elevates in some (but not all) subjects and is antecedent to hyperlipidemia. We caution that this example shows but one of many ways in which a covariate  $C$  could explain the bias seen in the stratified result in Example 1, and it was contrived solely to limit the numerical complexity of illustrating Points 2 and 3; realistic examples can become quite complex (see, for example, Robins<sup>2-7</sup>).

EXAMPLE 4

Suppose that the 12 types listed in Tables 1 and 2 can be subclassified into six subtypes according to the subject's  $C$  status when smoking and when quitting:

| Subtype | C status if |         |
|---------|-------------|---------|
|         | Smoker      | Quitter |
| 1       | 2           | 2       |
| 2       | 2           | 1       |
| 3       | 2           | 0       |
| 4       | 1           | 1       |
| 5       | 1           | 0       |
| 6       | 0           | 0       |

Note that quitting never elevates  $C$ . Let  $p_{ik}$  denote the proportion of Type  $i$  subjects of Subtype  $k$  (for example,  $p_{43}$  is the proportion of subjects for whom smoking produces cardiovascular disease via the hyperlipidemia pathway and for whom  $C = 2$  if they smoke but  $C = 0$  if they quit). Of the 72 possible  $p_{ik}$  ( $i = 0, \dots, 11, k = 1, \dots, 6$ ), assume all are zero except:

$$\begin{aligned}
 p_{ik} &= 0.004 \text{ for } i,k = 1,2; 1,4; 1,6; 2,3; 2,5; \\
 &0.016 \text{ for } i,k = 0,3; 0,5; 4,2; 4,4; 4,6; \\
 &0.020 \text{ for } i,k = 2,1; 6,3; 6,5; 10,2; 10,4; 10,6; \\
 &0.060 \text{ for } i,k = 9,3; 9,5; 11,3; 11,5; \text{ and} \\
 p_{0,1} &= 0.080, p_{6,1} = 0.100, p_{11,6} = 0.108, p_{6,6} = 0.252.
 \end{aligned}$$

The reader may verify that summing these numbers across the subtype index  $k$  gives back the distribution

of types in Example 1, for example,  $\sum_k p_{4k} = 0.048 = p_4$ , and that, as in Example 1, the only effects of continued smoking are mediated through hyperlipidemia (that is, since there are no type 3k, 5k, or 7k subjects, there are no direct effects of smoking).

If we randomize smoking cessation, the expected proportions within each smoking group will appear as in Table 6. Smoking appears protective in all but one CZ stratum; upon performing an ordinary stratified analysis across the six subtables in Table 6, we would obtain a risk difference (standardized to the total distribution) of  $-0.167$ . In this example, however, smoking is never preventive, and so the conventional approach is biased. The reason that the ordinary stratified estimate is biased is because exposure affects the level of C. For example, the risk of  $C = 2$  among smokers is  $0.20 + 0.20 = 0.40$ , whereas the risk of  $C = 2$  among quitters  $0.08 + 0.12 = 0.20$ .

Let  $I_{jmn}$  be the expected proportion in stratum level  $(m, n)$  (that is,  $C = m, Z = n$ ) when  $X = j$ . Let  $R_{jmn}$  be the expected disease incidence among subjects at level  $X = j, C = m, Z = n$  and let  $R_{jmnB}$  be the expected disease incidence among subjects at level  $X = j, C = m, Z = n$  if the effect of hyperlipidemia were blocked. Note  $R_{jm0} = R_{jm0B}$  since subjects in stratum  $(j, m, 0)$  are normolipidemic. Then indirect effects  $p_4 = \sum_k p_{4k}$  are given by

$$\sum_m I_{1m1}(R_{1m1} - R_{1m1B}) - \sum_m I_{0m1}(R_{0m1} - R_{0m1B}) \quad (5)$$

and direct effects are given by

$$\begin{aligned} (R_{1c} - R_{0c}) - [\sum_m I_{1m1}(R_{1m1} - R_{1m1B}) \\ - \sum_m I_{0m1}(R_{0m1} - R_{0m1B})] \\ = \sum_m \sum_n (I_{1mn} R_{1mnB} - I_{0mn} R_{0mnB}) \end{aligned} \quad (6)$$

This is a generalization of Eqs 3 and 4 to the stratified case.

Now, computing from the  $p_{ik}$  given earlier, one may directly verify that within levels of C and the exposure X, the counterfactual incidences  $R_{jm1B}$  equal the actual expected incidences  $R_{jm0}$  among normolipidemics; in other words, the exchangeability assumptions E1 and E2 hold within strata of C. Substitution of the numbers from Table 6 into Eq 5, under the assumption  $R_{jm1B} = R_{jm0}$ , thus yields the expected estimate of the indirect effect

$$\begin{aligned} p_4 &= 0.2(0.7 - 0.6) - 0.08(1 - 1) \\ &+ 0.1(0.4 - 0.2) - 0.008(1 - 0) \\ &+ 0.272(0.735 - 0) - 0.288(0.125 - 0.111) \\ &= 0.048, \end{aligned}$$

as it should. Consequently, the expected estimate of the direct effect is zero (since it equals the expected crude estimate of 0.048 minus the expected indirect effect estimate of 0.048). This illustrates how the G-computation algorithm can be used to control confounding of direct and indirect effects.

If, within joint strata of X and C, the risk of cardiovascular disease among controlled (blocked) hyperlipidemics equals the risk among normolipidemics (so

TABLE 6. Expected Proportions (Incidences) for Randomized Trial in Example 4

| Smoking Status | Level of C | Level of Z | Proportion at CZ Level mn Given X Level j ( $I_{jmn}$ ) | Cardiovascular Disease Incidence at CZ Level mn Given X Level j ( $R_{jmn}$ ) |
|----------------|------------|------------|---------------------------------------------------------|-------------------------------------------------------------------------------|
| X = 1          | C = 2      | Z = 1      | 0.200                                                   | 0.700                                                                         |
|                |            | 0          | 0.200                                                   | 0.600                                                                         |
|                |            | 1          | 0.100                                                   | 0.400                                                                         |
|                | 1          | 0          | 0.100                                                   | 0.200                                                                         |
|                |            | 1          | 0.272                                                   | 0.0735*                                                                       |
|                |            | 0          | 0.128                                                   | 0                                                                             |
|                |            |            |                                                         | $R_{1c} = 0.340$                                                              |
| 0              | 2          | 1          | 0.080                                                   | 1                                                                             |
|                |            | 0          | 0.120                                                   | 1                                                                             |
|                |            | 1          | 0.008                                                   | 1                                                                             |
|                | 1          | 0          | 0.072                                                   | 0                                                                             |
|                |            | 1          | 0.288                                                   | 0.125                                                                         |
|                |            | 0          | 0.432                                                   | 0.111†                                                                        |
|                |            |            |                                                         | $R_{0c} = 0.292$                                                              |

\* 5/68 exactly.

† 1/9 exactly.

that assumptions E1 and E2 hold within strata of C and thus  $R_{jm1B} = R_{jm0}$ , Eq 5 for indirect effects and Eq 6 for direct effects simplify to

$$\sum_m [I_{1m}(R_{1m} - R_{1m0}) - I_{0m}(R_{0m} - R_{0m0})] \quad (5')$$

and

$$\sum_m (I_{1m}R_{1m0} - I_{0m}R_{0m0}) \quad (6')$$

where  $I_{jm}$  is the expected proportion of subjects with  $X = j$  who have  $C = m$ , and  $R_{jm}$  is the expected rate of disease in subjects with  $X = j$  and  $C = m$ . The expected proportions in Eq 5' and Eq 6' can be unbiasedly estimated from the observed data by the corresponding sample proportions.

If, in addition, exposure had not affected the level of C (that is,  $I_{1m} = I_{0m} = I_m$ ), Eq 6' would reduce to  $\sum_m I_m(R_{1m0} - R_{0m0})$ , which is, among the normolipidemics, the expected standardized risk difference standardized to the distribution of C among the normolipidemics.

### 5. The Effect of Interactions

If exposure and the cofactor interacted (in the sense of Ref 10) to cause disease, then the G-computation algorithm Eqs 3 and 4 no longer give indirect and direct effects even in the randomized cofactor-intervention trial of Section 3. To see this, consider the following causal type.

- 12) Hyperlipidemia occurs unaffected by smoking, and then smoking and hyperlipidemia interact to cause disease (Statements 1-3 are true, Statements 4-6 are false).

Hyperlipidemia and cardiovascular disease occurrence for this type would be as given in the first row of Table 7. Note that Type 12 cannot represent an indirect effect of exposure since, by definition, there is no causal pathway from exposure to the cofactor (since the cofactor hyperlipidemia occurs whether or not a subject is exposed). Thus, the exposure effect in Type 12 subjects is direct. Suppose the whole population were composed of Type 12 individuals, and the randomized cofactor-intervention trial of Section 3 were carried out. In such a trial, Eq 3 equals 1. To see this, note that both  $I_{11} = 1$  and  $I_{01} = 1$ , since hyperlipidemia occurs irrespective of exposure. Furthermore,  $R_{11} = 1$  but  $R_{11B}$ ,  $R_{01}$ , and  $R_{01B}$  are 0, since disease does not occur except in hyperlipidemic smokers. Eq 3, therefore, no longer estimates the indirect exposure effects. Nonetheless, Eq 3 still estimates the smoking effect that could be eliminated by controlling (all) hyperlipidemia, and Eq 4 estimates the exposure effect that would remain once hyperlipidemia were controlled.

For Type 12 subjects, the smoking effect that could be eliminated by controlling hyperlipidemia does not equal the indirect effect of smoking, because the effect of smoking on cardiovascular disease can be blocked by controlling hyperlipidemia even though smoking does not cause hyperlipidemia.

When the smoking effect that could be eliminated by controlling hyperlipidemia differs from the indirect effect of smoking, it is the former effect that would be the parameter of public health interest whenever (1) there exists a public health intervention that controls hyperlipidemia (for example, prescription of a cholesterol-lowering drug or diet), but (2) there is no intervention that specifically blocks exposure's effect on elevating serum lipids. On the other hand, if there was no intervention available that could directly lower serum lipids, but there was a drug that would specifically block smoking's ability to elevate lipid levels, then it would be the indirect effects of smoking that would be the parameter of public health interest.

We have seen that the G-computation algorithm Eqs 3 and 4 can fail to estimate the indirect and direct effects of exposure even in the randomized cofactor intervention trial of Section 3. Is there some other computational formula that can estimate the direct and indirect effects in such a trial? We will show that there is no such formula. That is, direct and indirect effects are not identifiable in this trial. To do so, we require a more careful definition of direct and indirect effects. Specifically, we need to distinguish pure indirect effects from total indirect effects. Consider the following causal type:

- 13) Smoking would cause hyperlipidemia, and then smoking and hyperlipidemia would interact to cause disease (Statements 1 and 3 true; Statements 2 and 4-6 false).

Hyperlipidemia and cardiovascular disease occurrence for this type would be as given in Table 7. Disease in type 13 subjects is a result of both direct and indirect exposure effects. If, as we continue to assume, exposure never prevents disease or competes to cause disease, then Types 4 and 13 represent all possible indirect effects of exposure. Type 4 represents the pure indirect effects. Type 4 plus Type 13 represent the total indirect effects.

We now show that one may be unable to identify the total indirect effects from our randomized cofactor intervention trial. Consider two populations. The first population is equally divided between Types 8, 9, 12, and 13. Since Types 8 and 9 represent no exposure effect, Type 12 represents direct effects, and Type 13 represents indirect effects, we would have one-quarter

of this population with indirect effects. The second population is composed of only Types 9 and 12, each in equal proportion, so there are no indirect effects. Nonetheless, the two populations would produce exactly the same expected data in the randomized cofactor interaction trial of Section 3. The reader can check this using Tables 2 and 7. It follows that there can be no estimator that can identify the total indirect effects in this trial. If we define the pure direct effects to be the total exposure effects minus the total indirect exposure effects, then it immediately follows that we cannot identify the fraction of the total effects that are pure direct effects either.

Next, define the total direct effects to be the total (net) effect of exposure minus the fraction of pure indirect effects, that is, minus the fraction of Type 4 subjects. Note that Type 13 subjects contribute both to the total direct and total indirect effect, so that the sum of the total direct plus total indirect effect will exceed the total exposure effect if Type 13 subjects are present. We now show that we may be unable to identify (that is, separate) either the pure indirect effect of exposure or the total direct effect in our randomized cofactor intervention trial. Consider two populations. The first is equally divided among Types 4, 10, 13, and 14, where Type 14 is as defined in Table 7. The second population is composed of only Types 10 and 13 in equal proportion. The two populations would have the same expected outcomes in the randomized cofactor intervention trial, and yet only the first population has Type 4 subjects.

These nonidentifiability results are not so troubling when we again recall that the parameter of public health interest would usually be the smoking effect that could be eliminated by controlling hyperlipidemia,

that is,  $(p_4 + p_{12} + p_{13})$ . In our randomized cofactor intervention trial,  $p_4 + p_{12} + p_{13}$  can still be estimated by Eq 3 (see Robins<sup>1-7</sup>). Furthermore, even if there is no intervention on the cofactor, if we can find a covariate C such that E1 and E2 are satisfied within strata of C and the exposed and unexposed are exchangeable, then  $(p_4 + p_{12} + p_{13})$  can be estimated using the G-computation algorithm Eq 5 or Eq 5'.

We wish to caution the reader who plans to consult Refs 1-7 that the definition of the direct effect of exposure used in this paper differs from that in Refs 1-7. Specifically, what we have called the "exposure effect that would remain once the intermediate hyperlipidemia was controlled" was called in Refs 1-7 "the direct effect of exposure controlling for the intermediate Z when Z is fixed at (i.e., controlled to be) 0." What was called in Refs 1-7 "the direct effect of exposure controlling for the intermediate Z when Z is fixed at 1" is the exposure effect that would be observed if all study subjects were made hyperlipidemic.

It can be shown that we could separately identify the pure and total direct and indirect effects if we have data available from a crossover trial with no carryover effects in which both exposure and the cofactor intervention are randomly assigned in both time periods. Even without randomization, we could identify direct and indirect effects in such a crossover study provided both that data on appropriate confounders were available and that there were no carryover effects. Of course, we would rarely have data from such a study. Separation of direct from indirect effects requires a crossover study without carryover effects because of the need to differentiate those exposed hyperlipidemics whose hyperlipidemia is attributable to smoking from

TABLE 7. Occurrence of Hyperlipidemia and Cardiovascular Disease Events under Different Conditions for the Causal Types 12, 13, and 14

| Type | Statement Number (See Text) |                 |                                    |                |                 |                |
|------|-----------------------------|-----------------|------------------------------------|----------------|-----------------|----------------|
|      | 1                           | 2               | 3                                  | 4              | 5               | 6              |
|      | Serum Lipid Status If:      |                 | Cardiovascular Disease Status* If: |                |                 |                |
|      |                             |                 | Smoker (X = 1)                     |                | Quitter (X = 0) |                |
|      | Smoker (X = 1)              | Quitter (X = 0) | Hyper (Z = 1)                      | Normal (Z = 0) | Hyper (Z = 1)   | Normal (Z = 0) |
| 12   | 1                           | 1               | 1                                  | (0)            | 0               | (0)            |
| 13   | 1                           | 0               | 1                                  | (0)            | (0)             | 0              |
| 14   | 0                           | 0               | (1)                                | 0              | (0)             | 0              |

In body of table, 1 = statement true (event occurs), and 0 = statement false (event does not occur). Hyper = hyperlipidemia (Z = 1), Normal = normal serum lipids (Z = 0).

\* Outcomes in parentheses are counterfactual (and so remain unobserved) if there is no experimental manipulation of serum lipid status, even in a double-blind crossover trial of smoking.

those whose hyperlipidemia is not. That is, we need to estimate the joint distribution of the variable representing lipid status under exposure and the variable representing lipid status under nonexposure. To do so, it is necessary to observe the same subject both under exposure and nonexposure.

When the exposure and/or intermediate can also prevent disease or compete to cause disease, the total indirect effect of exposure is defined to be the expected proportion developing disease in the exposed minus the expected proportion in the exposed had exposure's effect on the intermediate  $Z$  been blocked (that is, had  $Z$  remained at its unexposed value); the total direct effect of exposure is the expected proportion developing disease in the exposed minus the expected proportion in the exposed when unexposed but with  $Z$  remaining at its exposed value. The pure direct (indirect) effect of exposure is the total exposure effect minus the total indirect (direct) effect.

### Conclusion

When exposure and the intermediate  $Z$  interact to cause disease (that is, there are subjects of Types 12, 13, and 14), we cannot hope to separate direct from indirect effects, with the possible exception of certain crossover studies with no carryover effects. Nonetheless, standard adjustment for or stratification on the intermediate will allow us to estimate the exposure effect that would remain after control of the intermediate, provided that the exposed and nonexposed are exchangeable and the additional exchangeability assumptions E1 and E2 hold.

Assumptions E1 and E2 may hold only within levels of a covariate (that is, confounder)  $C$  that occurs before the intermediate  $Z$  but subsequent to exposure. In this case, the exposure effect that could be eliminated by control of the intermediate can be estimated using Eq 5' if the exposed and unexposed are exchangeable as well. Furthermore, the exposure effect that would remain after control of the intermediate can be estimated using Eq 6'. In fact, Eqs 5' and 6' can be used validly to estimate these effects regardless of the mechanism by which exposure and the intermediate jointly affect disease. For instance, estimation validity is not compromised even if exposure, the intermediate, or both, prevent disease in some subjects.<sup>1-7</sup> In contrast, conventional simultaneous adjustment for both the confounder  $C$  and the intermediate will be biased if  $C$  is itself affected by exposure.

In the presence of interaction, it is the potentially estimable parameter—the fraction of the exposure effect that could be eliminated by control of the inter-

mediate—that usually will be the parameter of public health interest, and not the nonidentifiable parameter, the fraction attributable to the indirect effect of exposure.

Implicit in the very definition of the "causal effect of exposure" used in this paper is the assumption that if an event (for example, the presence of hyperlipidemia at  $t_1$ ) would have occurred both under exposure and nonexposure, then exposure was not a cause of the event. We made this assumption even though we recognize that the exposure could have affected the time at which hyperlipidemia first occurred and yet hyperlipidemia would be present at  $t_1$  regardless of exposure.<sup>11-14</sup> We conclude with several comments regarding this assumption. First, even if exposure causally influenced the time at which hyperlipidemia develops, it is perfectly logical to say that exposure had no effect on the dichotomous event, the presence or absence of hyperlipidemia at  $t_1$ . Nonetheless, we agree that, in principle, it is usually preferable to consider the causal effects of an exposure in terms of exposure's effect on the occurrence time of an event.<sup>11-14</sup> Robins<sup>2-7</sup> provides a rigorous treatment of direct and indirect effects based on a general causal model for the effect of an exposure on time to occurrence. In this paper, however, we chose to use a simple causal model with dichotomous intermediate and outcome variables, because the simple model retains the essential logical and philosophical details of the more general causal model, and yet avoids the mathematical complexities of the general model.

### Acknowledgments

We wish to thank Charles Poole, W. Douglas Thompson, Philip Kass, Malcolm Maclure, and Jennifer Kelsey for their many helpful suggestions regarding this presentation, and Warren Browner and Matthew Longnecker for suggesting improvements to the central example.

### References

1. Robins JM. The control of confounding by intermediate variables. *Stat Med* 1989;8:679-701.
2. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods: application to the healthy worker survivor effect. *Math Modeling* 1986;7:1393-152.
3. Robins JM. Addendum and errata for a new approach to causal inference in mortality studies with sustained exposure periods: application to the healthy worker survivor effect. *Comput Math Appl* 1987;14:923-953.
4. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chron Dis* 1987;40 (suppl):139S-161S.
5. Robins JM. The analysis of randomized and non-randomized

- AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. *Health Service Research Methodology: A Focus on AIDS*. NCHSR, U.S. Public Health Service, 1989;113-159.
6. Robins JM. Estimating the causal effect of a time-varying treatment on survival using a new class of failure time models. *Commun Stat* (in press).
  7. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun Stat* (in press).
  8. Greenland S, Robins JM. Identifiability, exchangeability and confounding. *Int J Epidemiol* 1986;15:413-419.
  9. Robins JM. Confidence intervals for causal parameters. *Stat Med* 1988;7:773-785.
  10. Greenland S, Poole C. Invariants and noninvariants in the concept of interdependent effects. *Scand J Work Environ Health* 1988;14:125-129.
  11. Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol* 1988;128:1185-1197.
  12. Robins JM, Greenland S. Estimability and estimation of excess and etiologic fractions. *Stat Med* 1989;8:845-859.
  13. Robins JM, Greenland S. The probability of causation under a stochastic model for individual risk. *Biometrics* 1989;45:1125-1138.
  14. Robins JM, Greenland S. Estimability and estimation of years of life lost due to a hazardous exposure. *Stat Med* 1991;10:79-93.
  15. Lewis DK. *Counterfactuals*. Cambridge: Harvard University Press, 1973.
  16. Holland PW. Reader reaction: confounding in epidemiologic studies. *Biometrics* 1989;45:1310-1316.
  17. Rubin DB. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Stat Sci* 1990;5:472-480.
  18. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat* 1978;6:34-58.