International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2011), 79, 3, 401–426 doi:10.1111/j.1751-5823.2011.00158.x

TECHNICAL REPORT
R-369
November 2011

Adjustments and their Consequences—Collapsibility Analysis using Graphical Models

Sander Greenland[1] and Judea Pearl[2]

[1]*Departments of Epidemiology and Statistics, University of California, Los Angeles, CA 90095-1772, USA*
E-mail: lesdomes@ucla.edu
[2]*Cognitive Systems Laboratory, Departments of Computer Science and Statistics, University of California, Los Angeles, CA 90095-1772, USA*
E-mail: judea@cs.ucla.edu

Summary

We review probabilistic and graphical rules for detecting situations in which a dependence of one variable on another is altered by adjusting for a third variable (i.e., non-collapsibility or non-invariance under adjustment), whether that dependence is causal or purely predictive. We focus on distinguishing situations in which adjustment will reduce, increase, or leave unchanged the degree of bias in an association that is taken to represent a causal effect of one variable on the other. We then consider situations in which adjustment may partially remove or introduce a bias in estimating causal effects, and some additional special cases useful for case-control studies, cohort studies with loss, and trials with non-compliance (non-adherence).

Key words: Bias; causality; causal models; collapsibility; compliance; confounding; graphical models; instrumental variables; mediation analysis; odds ratio.

1 Introduction

A common analysis question is whether adjustment for a variable C will reduce, increase, or leave unchanged the degree of association between two other variables, say X and Y. The question comes into focus at two stages of the analysis. First, the investigator may have a simple qualitative model of the data-generating process and may wish to test whether predictions of that model match the observed changes in associations that are induced by various adjustments. Second, when the association of interest is taken to represent the causal effect of X on Y, the investigator may wish to minimize bias by adjusting for the proper set of variables. In both stages, predicting the effect of an adjustment on a given measure of association becomes a question of central concern.

Many authors have considered aspects of adjustment effects using causal diagrams (e.g., Pearl, 1995; Greenland *et al.*, 1999a; Robins, 2001; Greenland, 2003; Chaudhuri & Richardson, 2003; Hernán *et al.*, 2004; Schistermann *et al.*, 2009; VanderWeele, 2009a; Didelez *et al.*, 2010). In this paper, we attempt to encompass these earlier results and reviews to provide a comprehensive

guide to when adjustment may not fully remove or may even introduce a source of bias, based on examining graphical models for the data-generating process. We contrast the impacts of adjustment on odds ratios versus other measures such as risk differences and mean differences. We show how adjustment may be used to discriminate among competing models. We also consider some additional special cases useful for estimating causal effects from case-control studies with differential selection, cohort studies with differential loss to follow-up, and trials with non-compliance (non-adherence).

Specifically, we focus on graphical (and hence qualitative) tools for recognizing situations in which an adjustment for C can or cannot alter a measure of the dependence of Y on X. These tools apply whether that dependence is causal (i.e., a comparison of Y distributions under different interventions on X) or purely predictive (i.e., a comparison of Y distributions in subpopulations defined by X). Central to our discussion will be concepts of *collapsibility*. When an adjustment does not alter a measure, the measure is said to be *collapsible* over C or *invariant* with respect to the adjustment. Conversely, if an adjustment alters a measure, the measure is said to be *non-collapsible* over C or *non-invariant* with respect to the adjustment. Thus, our paper reviews graphical conditions for collapsibility or invariance of measures with respect to particular adjustments.

When collapsibility cannot be deduced from the graph alone, we will say that we "expect" adjustment to change our estimate (i.e., we expect non-collapsibility). That is because in those cases, collapsibility will require sharp (dimension-reducing) constraints on the joint distribution of the variables in the graph, and these constraints are often unnatural in form (e.g., requiring perfect parametric cancelations; see Whittemore, 1978). Otherwise, we will use the phrase "we might ordinarily expect" to indicate less precisely those situations we expect to hold in ordinary applications in the health and social sciences, when violations seem to require unusual distributions. We will draw heavily on statistical results for collapsibility of risk, rate, and odds-based measures in contingency tables and binary regression (e.g., Whittemore, 1978; Samuels, 1981; Ducharme & LePage, 1986; Gail, 1986; Wermuth, 1987; Greenland & Mickey, 1988; Geng, 1992; Frydenberg, 1990; Clogg *et al.*, 1995; Guo & Geng, 1995; Greenland, 1996; Geng & Li, 2002; Geng *et al.*, 2002; Janes *et al.*, 2010). Nonetheless, as with earlier graphical articles our discussion applies to continuous variables as well, due to the non-parametric nature of the formulas and graphical results we employ.

The paper begins by briefly reviewing necessary concepts and results from probability, population sciences, and graph theory in the general form we will need. We then explain, in a series of simple examples, how these concepts can be used to determine the qualitative impact of various adjustments on bias in estimating causal effects. The examples provide the basic graphical structures (subgraphs) that frequently appear within more complex models, and are useful because bias arising within a subgraph would usually arise within any embedding model.

2 Conditioning, Summarization, and Standardization

Unless stated otherwise, all subpopulations and distributions we discuss will be within the source population of the study, by which we mean the population serving as the source of study subjects (not person-time). By *conditioning on* a variable (or set of variables) C we will mean examining relations within levels of C (i.e., within strata defined by single values of C). By *summarization* over C we will mean summarization of conditional (C-specific) dependencies across C. This definition includes pure conditioning, in which the summary is the list (vector) of C-conditional dependence measures, such as C-specific risk differences, risk ratios, log odds ratios, and so on; it also includes averaging these measures over C.

In practice, summarization is usually done using a regression coefficient under a highly fictional model in which the coefficient relating X to Y is assumed constant across C (known as homogeneity, uniformity, parallelism, "no interaction," or "no effect modification"). Any average coefficient must then equal this constant, so the method of averaging does not matter. We will, however, focus on the general case, free of homogeneity assumptions, in which the averaging method can be important.

By *adjustment for C* we will then mean one of the many ways one might adjust the dependence of Y on X to account for the relations of C to Y and X. This definition includes both conditioning and averaging of C-specific (conditional) measures, but also includes standardization (comparisons of average outcomes), which can diverge from averaging of C-specific measures. In common usage, the word "control" is often used as a synonym for "adjustment" and will be used that way here. Adjustment should not however be confused with experimental control (manipulation or intervention) as the consequences of experimental control of C can be quite different from conditioning and other adjustments for C; we will point out some parallels and divergences when discussing adjustments for causes of C.

Of special focus will be comparisons across X of the distribution of Y given X and C, $p(y|x, c)$, when this distribution is averaged over a specific distribution $p^*(c)$ for C. Following common usage in demography, epidemiology, and survey research, when $p(y|x, c)$ is derived from population observations, we will call the joint distribution $p(y, x, c)$ for these observations the *source-population* distribution, and $p^*(c)$ the *standard* distribution, which need not equal the marginal distribution $p(c)$ of C in the source population. We will denote the resulting averages by

$$p\{y|x; p^*(c)\} \equiv \sum_c p(y|x, c)p^*(c) \tag{1}$$

(the sum is over all values of C). We will assume that $p^*(c) = 0$, whenever $p(x, c) = 0$ so that the average remains defined. Such averages are commonly known as the "probability of $Y = y$ given $X = x$, *standardized to* (averaged over) $p^*(c)$"; they are the expectations of the indicator of the event $Y = y$ in a population which the conditional distributions of Y given X and C are $p(y|x, c)$, X is constant at level x, and the marginal distribution of C is $p^*(c)$.

When Y is a health or disease outcome variable, $p\{y|x; p^*(c)\}$ is the classical *standardized risk* of $Y = y$ when $X = x$ and $p^*(c)$ is the standard distribution (Rothman *et al.*, 2008, ch. 3 and 21). In particular, when $p^*(c) = p(c)$,

$$p\{y|x; p(c)\} \equiv \sum_c p(y|x, c)p(c) \tag{2}$$

are the risks of $Y = y$ standardized to the total source population. Also of traditional interest are averages over a specific "reference standard" $p^*(c) = p(c|x_r)$, where x_r is a contextually determined reference value of X (e.g., exposure or non-exposure):

$$p\{y|x; p(c|x_r)\} \equiv \sum_c p(y|x, c)p(c|x_r). \tag{3}$$

Note that when (3) is evaluated at $X = x_r$, C disappears from the expression:

$$p\{y|x_r; p(c|x_r)\} = \sum_c p(y|x_r, c)p(c|x_r) = p(y|x_r).$$

A fundamental research question is how much variation in Y can be attributed to variation in X rather than C. A non-parametric answer can be obtained by examining how averages, such as (1), (2), or (3) vary with X. The resulting comparisons across X are called C-standardized measures of the dependence of Y on X. The C distribution $p^*(c)$ is held constant across these comparisons,

thus removing this distribution as a factor contributing to variation in the Y distribution across X. When the standard (weighting) distribution $p^*(c)$ and the Y dependence $p(y|x, c)$ are derived from the same population, as in (2) and (3), the resulting average is said to be *population standardized.* Examples include the *standardized morbidity ratio* (SMR) for subpopulation with $X = x_r$, which divides (3) evaluated at $X = x_r$ by (3) evaluated at another value of X; it simplifies to $p(y|x_r)/p\{y|x; p(c|x_r)\}$ (Rothman *et al.*, 2008, ch. 4).

Standardized distributions are equivalent to the distribution of Y given X obtained after inverse-probability weighting of the joint distribution using the distribution of X given C (Robins *et al.*, 2000; Sato & Matsuyama, 2003; Rothman *et al.*, 2008, ch. 21). For example, $p(y|x, c)p(c) = p(y, x, c)/p(x|c)$ and so

$$p\{y|x; p(c)\} = \sum_c \frac{p(y, x, c)}{p(x|c)},$$

More generally,

$$p(y|x, c)p^*(c) = \frac{p(y, x, c)p^*(c)}{p(x|c)p(c)}$$

and so

$$p(y|x; p^*(c)\} = \sum_c \frac{p(y, x, c)r^*(c)}{p(x|c)}$$

where $r^*(c) = p^*(c)/p(c)$ measures divergence of the chosen standard $p^*(c)$ from the source-population standard $p(c)$. *Marginal structural modeling* (Robins *et al.*, 2000; Rothman *et al.*, 2008, ch. 21) directly models these quantities, typically using smoothed estimates of $p(x|c)$. When C is sufficient for confounding control (see further), the contrasts among these quantities (e.g., differences and ratios) are then referred to as marginal effects.

3 Measure Averaging and Collapsibility

A common question in practice is when C can be ignored in an analysis, for example, when does $p\{y|x; p^*(c)\} = p(y|x)$ so that the standardization step can be skipped? The following basic collapsibility results provide some answers and have been noted in various forms at least since Yule (1934):

(a) Any standardized probability (1) will simplify to $p(y|x)$ if C is independent of Y given X, that is, if $p(y|x, c) = p(y|x)$ then $p\{y|x; p^*(c)\} = p(y|x)$.
(b) Population-standardized probabilities (2) and (3) will simplify to $p(y|x)$ if C and X are marginally (unconditionally) independent; that is, if $p(c, x) = p(c)p(x)$ then

$$p\{y|x; p(c|x_r)\} = p\{y|x; p(c)\} = p(y|x).$$

Result (a) follows from the fact that if $p(y|x, c) = p(y|x)$ then $p(y|x)$ factorizes out of the summation over c and the latter summation becomes 1. Result (b) follows by noting that if $p(c, x) = p(c)p(x)$ then expression (2) becomes

$$\sum_c \frac{p(y, x, c)p(c)}{p(x)p(c)} = \sum_c \frac{p(y, x, c)}{p(x)} = \frac{p(y, x)}{p(x)} = p(y|x),$$

and since $p(c|x_r) = p(c)$, expressions (2) and (3) are equal. It follows from these results that population-standardized measures (such as differences and ratios of population-standardized probabilities) are collapsible over C if either (a) C is independent of Y given X, or (b) C and X are marginally independent.

Standardized measures are constructed by taking averages over C *before* comparisons (e.g., ratios or differences) across X. Many other familiar adjusted measures are instead derivable by taking averages of comparisons within levels of C; that is, they average over conditional measures of association, *after* comparison across X. Examples include inverse-variance (information)-weighted averages. Recalling Jensen's inequality (an average of a non-linear function does not equal the function applied to the averages), it should not be surprising to find divergences between collapsibility conditions depending on the step at which averaging is done (Samuels, 1981, sec. 3).

Standardized difference and ratio measures can be rewritten as averages of conditional measures. For example, in comparing two levels x_1 and x_0 of X using formula (2), the difference is

$$\sum_c p(y|x_1, c)p(c) - \sum_c p(y|x_0, c)p(c) = \sum_c \{p(y|x_1, c) - p(y|x_0, c)\}p(c)$$

that weights the C-specific differences $p(y|x_1, c) - p(y|x_0, c)$ by $p(c)$. The standardized ratio is

$$\frac{\sum_c p(y|x_1, c)p(c)}{\sum_c p(y|x_0, c)p(c)} = \frac{\sum_c \{p(y|x_1, c)/p(y|x_0, c)\}w_0(c)}{\sum_c w_0(c)}$$

that weights the C-specific ratios $p(y|x_1, c)/p(y|x_0, c)$ by $w_0(c) = p(y|x_0, c)p(c)$ (Rothman *et al.*, 2008, p. 267). This ratio must fall within the range of the C-specific ratios. The same is true of other averages such as Mantel–Haenszel risk ratios (Rothman *et al.*, 2008, p. 275) and geometric mean ratios (such as those based on information weighting of log risk ratios).

3.1 Collapsibility Over Constant Measures

Even if C and X are marginally independent, not all averages of C-specific measures will be collapsible. Suppose, however, the C-specific measures are constant across C. If the unconditional measure equals this constant value, the measure is said to be *strictly collapsible* or *simply collapsible* over C (Whittemore, 1978; Ducharme & LePage, 1986; Geng, 1992). Since all averages of the C-specific measures must equal this constant, simple collapsibility implies collapsibility of these averages.

As an example, odds ratios are simply collapsible if X is independent of C given Y, as can be seen from the familiar XY "inversion" (symmetry) property of odds ratios: The C-specific odds ratios are

$$\frac{p(y_1|x_1, c)p(y_0|x_0, c)}{p(y_1|x_0, c)p(y_0|x_1, c)} = \frac{p(x_1|y_1, c)p(x_0|y_0, c)}{p(x_1|y_0, c)p(x_0|y_1, c)}.$$

If C is independent of X given Y the latter term becomes

$$\frac{p(x_1|y_1)p(x_0|y_0)}{p(x_1|y_0)p(x_0|y_1)} = \frac{p(y_1|x_1)p(y_0|x_0)}{p(y_1|x_0)p(y_0|x_1)},$$

thus, demonstrating simple collapsibility over C.

3.2 Non-collapsibility of Odds and Rate Measures

Because standardized risk differences and risk ratios are averages of C-specific values, simple collapsibility of risk differences and risk ratios implies collapsibility of standardized risk differences and risk ratios. Simple collapsibility of odds ratios and differences does not however imply collapsibility of the standardized odds ratios and differences. Instead, rather paradoxically, if C and Y are dependent given X but C is marginally independent of X, all population-standardized odds ratios will be collapsible but simple collapsibility cannot hold (Miettinen & Cook, 1981; Greenland *et al.*, 1999b). Parallel results hold for odds differences, as well as for rate ratios and differences (Greenland, 1996).

More generally, if an unadjusted (unconditional) measure is outside the range of the C-conditional measures, then the measure cannot be collapsible with respect to any average of C-conditional measures (such as a standardized risk difference or risk ratio). Nonetheless, an odds ratio or rate ratio may still be collapsible with respect to other adjustments. For example, standardized odds ratios constructed from (1) to (3) usually do not reduce to weighted averages of C-specific odds ratios. Thus, an odds ratio may be collapsible with respect to a particular standardization, yet may be non-collapsible with respect to any average over the C-specific odds ratios. This conflict complicates the interpretation of odds ratios and has led to much confusion in the literature. In particular, non-collapsibility over C with respect to averaging over odds ratios (or their logs) is sometimes called a "bias," but if C is sufficient for confounding control (see further), it does not correspond to a bias in estimating causal effects (Greenland *et al.*, 1999b).

3.3 Complete Collapsibility

Now suppose that conditioning on C does not alter the dependence of Y on X, that is, $p(y|x, c) = p(y|x)$ for all c and x (conditional independence of Y and C given X). Then conditioning on C cannot change *any* measure of dependence of Y on X, and any reasonable adjustment for C (whether standardization of probabilities or averaging of measures across levels of C) must produce a measure equal to the unconditional (unadjusted, marginal) measure. In other words, independence of C and Y given X implies simple collapsibility for *all* dependence measures. We will call this condition *complete collapsibility* over C: neither standardization nor conditioning nor averaging measures over C will change the dependence of Y on X. Complete collapsibility thus corresponds to a situation in which adjustments for C have no impact on bias for any measure.

3.4 Conditional Collapsibility

All the above definitions and concepts can be applied conditional on a set S of further covariates. For example, independence of Y and C given X and S implies *complete collapsibility given S* (after conditioning on S, further conditioning or adjustment for C does not change the dependence of Y on X given S). Similarly, an adjusted measure adjusted for C and S is *collapsible over C given S* if it equals its counterpart from adjusting for S only.

4 Connectivity and Associations in Directed Acyclic Graphs

There are now many introductory reviews of causal and probabilistic analysis using directed acyclic graphs (DAGs) (e.g., Greenland *et al.*, 1999a; Glymour & Greenland, 2008; Greenland & Pearl, 2010; Pearl, 2010a), as well as much more in-depth treatments (e.g., Pearl, 2009; Spirtes

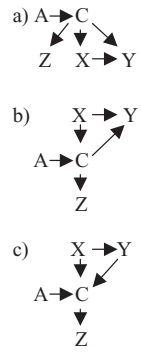

Figure 1. *Graphs with C connected to X and Y under all conditions.*

et al., 2001). Figure 1 gives three basic cases: In (a) the covariate C is a cause of X and Y, in (b) it is a mediator between X and Y, and in (c) it is an effect of both X and Y. We summarize briefly the graphical concepts we will use to analyze them. Throughout, we will assume the graph represents relations in a specific population under study. The results we will use apply even if the graph represents only conditional independencies rather than causal relations, although as in decision analysis their interest here derives from their causal interpretation (Kiiveri *et al.*, 1984; Spiegelhalter, 1987). Readers familiar with DAGs may skip this section.

Two *arrows* in a DAG are adjacent if each touches the same variable (whether by head or tail). A *path* between X and Y is a sequence of adjacent arrows going through the DAG from X to Y. A variable on a path between X and Y is an *interceptor* on the path. An interceptor C where two arrowheads meet (two arrows collide, as in the path from X to Y in Figure 1(c)) is a *collider* on the path, and the path is said to be *unconditionally blocked* at C. If instead C is where an arrowhead meets a tail (as in the path from X to Y in Figure 1(b)) it is a *mediator* on the path. Finally, if C is where two arrowtails meet (as in the path from X to Y in Figure 1(a)) it is a *fork* on the path. Note that all three of these conditions are only relative to a path; for example, in Figure 1(a), C is a fork on the path $X \leftarrow C \rightarrow Y$, a mediator on the path $A \rightarrow C \rightarrow Y$, and a collider on the path $A \rightarrow C \leftarrow X$; thus, it is not meaningful to speak of a variable as a mediator or collider without reference to the path on which it is so.

A path is said to be *unconditionally closed* or blocked at every collider and *unconditionally open* at every mediator or fork. Thus, a path is *unconditionally open* if it contains no collider; conversely, if the path contains a collider it is *unconditionally closed* or blocked. Two variables in a DAG are said to be unconditionally *d-connected* if there is an open path between them, and are unconditionally *d-separated* (Pearl, 1988, 1995, 2009) if there is no such path. The "*d*-" in these definitions stands for "directionally" and distinguishes these conditions from other concepts of separation. Nonetheless, because the popular DAG literature uses only directional concepts, in what follows we will shorten "*d*-connected" to "connected" and "*d*-separated" to "separated" (as in Greenland *et al.*, 1999a). We may then say that two variables are connected by all the open paths between them.

A path is *directed* if it contains only mediators (so that one moves from arrowhead to arrowtail at each variable in the path). If there is a directed path from one variable to another, the tail-end variable (the start) is called an *ancestor* of the variable at the ending arrowhead; and the variable at the final arrowhead (the end) is called a *descendant* of the starting variable. In a DAG, no variable is its own ancestor (i.e., there are no feedback loops). If the DAG is taken as a causal model, a variable is said to causally affect its descendants and be causally affected by its ancestors.

A path is *open given a set of variables S* if (i) *S* contains no mediator or fork on the path and (ii) any collider on the path is either in *S* or has a descendant in *S*; otherwise it is *closed given S* or *blocked by S*. Two variables in a DAG are *connected given S* if there is a path between them that is open given *S*; otherwise they are *separated given S*. Two *variables* are adjacent if they have an arrow between them. The variable at the tail is called the *parent* of the variable at the head, which is called the *child* of the tail variable. The set of parents of a variable X in a given DAG is denoted $\text{pa}(X)$. If X has no parent in the DAG, as in Figure 1(b), $\text{pa}(X)$ is empty and X is said to be *exogenous* in the DAG; otherwise X is *endogenous*, as in Figure 1(a) where $\text{pa}(X) = \{C\}$.

4.1 Compatibility between DAGs and Distributions

A distribution p and a DAG over a set of variables are said to be *compatible* if p factorizes into $\prod p(x|\text{pa}(X))$, where the product is over all the variables (this product is called the *Markov factorization* implied by the DAG). It can be shown (Pearl, 1988, 2009; Lauritzen, 1996; Spirtes *et al.*, 2001) that for any compatible p, two variables X and Y in a DAG will be independent given another set of variables S in the DAG if X and Y are separated by S. The converse is not true in general, but if S does not separate X and Y we would expect X and Y to be associated given S, because exceptions in which connected variables are nonetheless independent in p given S correspond to sharp constraints on p (and hence are sometimes referred to as "unstable" or "unfaithful" properties of the distribution; Pearl, 2009; Spirtes *et al.*, 2001).

The remainder of this paper is concerned primarily with describing properties of distributions compatible with a given graph. As a simple example of such a property, adjacent variables will always be connected and hence cannot be assumed independent, no matter what information we obtain about the remaining variables in the DAG. In other words, adjacent variables may remain dependent at any level of conditioning on the remaining variables in the DAG. Conversely, two non-adjacent variables X and Y are separated by at least one of $\text{pa}(Y)$ and $\text{pa}(X)$, and will be unconditionally independent if neither is a descendant of the other or a shared ancestor.

We emphasize that the statistical assumptions encoded by the DAG are carried by the *absence* of arrows: Given non-adjacent variables X and Y with X a non-descendant of Y, X and Y will be independent given the parents of Y. Thus, one should pay special attention to the contextual justification for leaving certain variables non-adjacent (Greenland, 2010). In light of this caution, in what follows we will assume each variable in a graph represents either a single variable or else a set of variables such that every arrow from or into the variable represents a set of arrows from or into every variable in the set it represents. If there are non-adjancies within the variables in the set, there will be fewer open paths for bias transmission than if there were none, and the results we present can be applied within the set to find more opportunities for bias control.

5 Separation and Collapsibility

We now consider when the relations encoded in a DAG signal the absence or presence of collapsibility. Considering the DAG as a probabilistic influence network or *Bayes net* (Lauritzen & Spiegelhalter, 1988; Pearl, 1986, 1988), information can flow from one point to another along open paths. In particular, if two variables are connected, then information can flow between them. This means we should not assume that connected variables are independent; in particular, new information obtained about a variable C may (upon conditioning on that information) alter our probabilities regarding any variable connected to it. Furthermore, if a variable C is connected to both X and Y, we should not be surprised if obtaining and conditioning on information about C alters the connection between X and Y.

Conversely, because separation implies independence in compatible distributions, we obtain the following two criteria for detecting collapsibility in a distribution given a compatible graph. These criteria allow us to recognize graphically when a dependency is invariant under conditioning and standardization:

(a) If C is separated from Y given X, then the dependence of Y on X will be completely collapsible over C (i.e., unaltered by adjustment for C).
(b) If C is separated from X unconditionally, then population-standardized measures of dependence of Y on X will be collapsible over C.

Both these criteria also apply conditional on a set S of covariates, and with C replaced by a set of covariates.

When comparing two levels x_1 and x_0 of X, criterion (a) applies to standardized differences and ratios of probabilities, such as $p\{y|x_1; p(c)\} - p\{y|x_0; p(c)\}$ and $p\{y|x_1; p(c)\}/p\{y|x_0; p(c)\}$ derived from expression (2). Under criterion (b) (unconditional separation of C and X), both these measures equal the analogous unconditional (unadjusted) measures $p(y|x_1) - p(y|x_0)$ and $p(y|x_1)/p(y|x_0)$ obtained by dropping $p(c)$ from the expressions. (These measures usually take Y to be a binary disease indicator with y denoting disease; our results apply to any Y and y.) Both criteria also apply when comparing two levels y_1 and y_0 of the outcome Y via standardized odds such as $p\{y_1|x; p(c)\}/p\{y_0|x; p(c)\}$, as well as to differences and ratios of these odds: Under either criterion, the resulting measures will be unchanged by standardization.

Each of criteria (a) and (b) is sufficient alone, but neither is necessary and so the converse of each is not quite correct. Nonetheless, if C is connected to Y conditional on X, then without more restrictions we will not have complete collapsibility for the dependence of Y on X; in particular, we would expect the C-conditional risk differences and risk ratios and their summaries to differ from the corresponding unconditional risk differences and risk ratios. Furthermore, if C is also connected to X unconditionally, we would expect non-collapsibility for averages across C of the risk differences and risk ratios. Again, collapsible exceptions require sharp constraints (unstable or unfaithful properties) for the distribution, and do not arise with binary C (Whittemore, 1978).

Whether the changes upon conditioning on C or adjustment for C represent increased or decreased bias depends upon further details, especially on the effect targeted for estimation (Glymour & Greenland, 2008; VanderWeele, 2009a). Intuitively, we might think conditioning on C will remove bias for estimating any effect of X on Y in Figure 1(a) and direct effects in Figure 1(b), but will create bias for estimating the total effect (the net of all effects transmitted through directed paths from X to Y) in Figure 1(b) and any effect in Figure 1(c). As discussed below, these intuitions are correct when targeting total-population effects.

When we consider odds $p(y_1|x, c)/p(y_0|x, c)$ and their comparisons conditional on C, instead of those computed from standardized probabilities, criterion (b) is no longer relevant. In its place, we have

(c) If C is separated from X conditional on Y, then the odds ratio will be collapsible over C.

As a partial converse, if C is connected to X conditional on Y and is connected to Y conditional on X, then we expect non-collapsibility over the C-conditional odds ratios (Didelez *et al.*, 2010). Again, we say "expect" because of special exceptions when C is not binary (Whittemore, 1978) and caution that, even in Figure 1(a), odds-ratio non-collapsibility partly represents a mathematical peculiarity of odds ratios and differences rather than pure confounding (Greenland *et al.*, 1999b). Parallel remarks apply to differences and ratios of rates (hazards) (Greenland, 1996), with Y now understood to contain both time at risk and the outcome indicator.

6 Relations to Causal Effects

The collapsibility results we have described do not assume the quantities at issue are related to causal effects. Historically, most statistical treatments of collapsibility left that relation informal, a practice that leads to pitfalls and thus has been gradually giving way to formal causal modeling. We thus now consider some formal connections and how they serve as guides to when adjustment will aid effect estimation.

6.1 Causal Models

Throughout, we assume the target of estimation is a *causal parameter* defined with respect to an explicit causal model. This target parameter is usually called "the" effect of X on Y, although there may be several different causal parameters (e.g., differences, ratios, etc.). A key question in causal analyses is whether the targeted effect equals a particular association measure before or after conditioning on a particular variable or set of variables. The answer will largely depend on the structure encoded by the causal model. The discrepancy between the targeted effect and an association measure is usually called the bias in taking the latter measure to equal the targeted effect; it is this bias which will concern us in what follows. When it is present, it persists asymptotically; we will not consider other forms of bias (such as small-sample and sparse-data bias).

There are now many formal causal models in widespread use, reviewed for example, in Pearl (2009). Fortunately, for our purposes all these models have extensive parallels and lead to the same graphical algorithms for bias identification and control. One approach defines $p(y|\text{do}[x])$ as the distribution Y would have upon intervening to set X to the value x for everyone in the population, when that is possible; a *causal model* on a set of variables V then specifies or constrains these intervention distributions for every subset X and Y of V. This do[x] formalism is closely related to the potential-outcome (counterfactual) model of causation, in which each individual is presumed to have a well-defined potential-outcome variable Y_x when administered level x of X, whether or not x is the level actually administered; in that case $p(y|\text{do}[x]) = p(y_x)$ (Pearl, 2009, ch. 7; Robins & Richardson, 2010).

In either formalism, care is needed in choice of X in order for the setting of X to a level x represented by do[x] or Y_x to make sense (Greenland, 2005a; Hernán, 2005). This would be so if X indicated a treatment such as a vaccination, but not if X indicated a defining property of an individual such as their sex. This issue is often subsumed under the topic of consistency (that $Y = Y_x$ when $X = x$; see Cole & Frangakis, 2009; VanderWeele, 2009b; Pearl, 2010b). It must also be assumed that any ancestor (cause) shared by two variables in the graph is in the graph (often called the "no unmeasured confounders" assumption, although this is a misnomer since variables in the graph may be unmeasured).

6.2 Sufficiency for Effect Estimation

When do [x] is well defined, we say a set of covariates S is *sufficient* or *admissible* for estimating total-population effects of X on Y if $p\{y|x; p(s)\} = p(y|\text{do}[x])$; that is, S is sufficient precisely in the case that standardization by $p(s)$ yields the effect of setting $X = x$. This condition is also known as sufficiency for confounding control, or adjustment sufficiency. When S is sufficient, $p\{y|x; p(s)\}$ is said to be *unconfounded*. S is *minimal sufficient* if it is sufficient but no subset of S is. Note that it is possible for some supersets of S to be insufficient even though S is sufficient and the superset contains no descendant of X or Y. A simple example is the "M diagram" $X \leftarrow A \rightarrow C \leftarrow B \rightarrow Y$ in which the empty set is sufficient, that is, $p(y|x) = p(y|\text{do}[x])$ but

$p\{y|x; p(c)\} \neq p(y|\text{do}[x])$ for most compatible distributions, even though C is a non-descendant of X and Y (the exceptions being "unfaithful," i.e., obey sharp constraints).

If S is sufficient for the total-population effects and (as elsewhere in the current paper) we assume no contagion (i.e., disease in one person does not cause disease in another person), we expect standardization by $p(s|x_r)$ to be sufficient for estimating effects in the subpopulation with $X = x_r$; that is, we expect $p\{y|x; p(s|x_r)\}$ to equal the effect of having set this subpopulation to $X = x$ instead of its actual setting of $X = x_r$ (Shpitser & Pearl, 2009; see also Robins & Richardson, 2010, appendix A). The converse is not correct: A set may be sufficient for some choices for x_r (some subpopulations) but insufficient for other choices of x_r; more generally, a set may be sufficient for some subpopulation effects but insufficient for others or for total-population effects (Joffe *et al.*, 2010). Contagion further complicates analysis of subgroups because then the distribution in one subgroup may depend on the distribution and hence alteration of other subgroups (Halloran & Struchiner, 1995).

6.3 Causal Graphs

In a DAG which represents a causal model on a set of variables V (i.e., in a *causal graph*), arrow directions always reflect temporal ordering, and for any two disjoint subsets of variables, X and Y in V:

(1) Every intervention distribution $p(y|\text{do}[x])$ from the causal model is compatible with the graph; that is, every separation in the graph remains a valid conditional independence in the post-intervention distribution $p(y|\text{do}[x])$, so that interventions destroy but never create dependencies.
(2) The conditional distribution $p(x|\text{pa}[X])$ remains invariant to interventions not directly affecting X ; that is, for any subset Z of V that excludes X and its parents,

$$p(x|\text{pa}[X], \text{do}[z]) = p(x|\text{pa}[X]).$$

It can be shown (Bareinboim *et al.*, 2011) that properties (1) and (2) are equivalent to the more usual definition of causal graphs, in which an intervention $\text{do}(x)$ removes all arrows pointing to X and sets X to a constant $X = x$ (Spirtes *et al.*, 2001; Pearl, 2009).
Two useful properties following from (1) and (2) are:

(1) Conditional on a set S sufficient for estimating the total population effect of X on Y, X will be connected to Y given S only via directed paths from X to Y.
(2) The set of parents of X, $\text{pa}[X]$, is a sufficient set for estimating the total-population effects of interventions on X, that is, $\text{p}\{y|x; p(s)\} = p(y|\text{do}[x])$ whenever $S = \text{pa}[X]$.

(Pearl, 2009, p. 24). Sufficient sets that do not contain $\text{pa}[X]$ can be recognized using the following criterion, based on paths that go through $\text{pa}[X]$. Specifically, a path from X to Y is said to be *back-door* (relative to X) if it starts with an arrow into X. A subset S of V then satisfies the *back-door criterion* for estimating the effect of X on Y if it (i) contains no descendant of X and (ii) blocks every back-door path from X to Y. Such a set is sufficient for effect estimation under the model (Pearl, 1995, 2009). Generalizations of the back-door criterion to sufficient sets containing descendants of X are available (Pearl, 2009, pp. 339–340; Shpitser *et al.*, 2010).

Now suppose that R is a sufficient set and T is a subset of R. Then, $S = R - T$ will also be sufficient if T is either independent of X given S or independent of Y given S and X. In

graphical terms, this result implies that if S separates T from either X or Y, then S as well as R will be sufficient. More generally, S will be sufficient if $T = T_1 \cup T_2$ with T_1 independent of X given S and T_2 independent of Y given S, X, and T_1 (see Robins, 1997, corollary 4.1). In graphical terms, this result implies that S will be sufficient if S separates T_1 from X and S, T_1, X together separate T_2 from Y. When R is not sufficient, the preceding condition still guarantees that standardizing by S using $p(s)$ will yield the same parameter as standardizing by R using $p(r)$ (Pearl & Paz, 2010, Thm. 3).

Tian *et al.* (1998) provide fast algorithms for finding minimally sufficient sets (when they exist) of observed variables in DAGs which may include unobserved variables. Suppose then we have found several minimal sufficient sets $S_1, \ldots, S_n$. Their union $S = \cup_i S_i$ is then also sufficient (Pearl & Paz, 2010, appendix), and efficiency considerations can lead to preference for the unique minimally sufficient subset that, given X, separates Y from all other members of S (Pearl 2011, section 3).

7 Effects of Conditioning and Adjustment in Basic Structures

We now apply the above results to determine the effects of conditioning and adjustment in basic graphical structures (subgraphs) that frequently appear within causal models for studies. We emphasize again that understanding simple structures is key to bias recognition in more complex graphs: A bias arising within a simple graph will typically arise in any supergraph containing (embedding) that simple graph. Conversely, we caution that the absence of bias within a simple graph does not imply or even suggest absence of bias in an embedding supergraph.

The examples are organized around graphical properties of variables, rather than around type of bias (which can only be determined after one writes down the graph or its algebraic analog). This organization is based on the following application guideline: First, one should write down the graph for the data generating process (or several, if the structure is in doubt and sensitivity analysis is needed), which is essentially a non-parametric representation of the data distribution function. Second, one should examine the graph to see the bias implications of various adjustments.

7.1 Effects of Conditioning on an Interceptor

In each graph in Figure 1, C intercepts a path from X to Y. In particular, C is connected to X unconditionally and conditional on Y, and to Y conditional on X. Thus, we expect non-collapsibility over C (non-invariance under C conditioning) for all measures; that is, in almost all cases, conditioning on C will change the risk differences, risk ratios, and odds ratios relating X to Y. Nonetheless, the meaning of this change with respect to bias is quite different across the graphs.

In Figure 1(a), the path between X and Y via C ($X \leftarrow C \rightarrow Y$) has a fork at C, and so is an open path; hence X and Y may be associated via this path. The association transmitted along this path is a source of bias for estimating the effect of X on Y, sometimes called "classical confounding" because the path contains a shared cause of X and Y (Greenland, 2003). Note that the key source of this confounding is that C is *uncontrolled*, not that it is unmeasured. For example, C may have been measured but left uncontrolled because it failed to have a "statistically significant" association with X or with Y. Conversely, C may have been controlled without being measured by virtue of design features (e.g., for practical purposes, a population-based study in Finland will have controlled for conventional American "race" categories by being almost entirely white, as well as other homogeneous Finnish-population characteristics).

Conditioning on C will block (close) the confounding path in Figure 1(a); hence if X has no effect on Y, then X and Y will be independent given C (independent within every stratum or level of C), reflecting correctly this absence of effect. Put more generally, C alone satisfies the back-door criterion and thus is sufficient for estimating effects; furthermore, it is minimal sufficient. Hence, to estimate an effect of X on Y, we should condition on C. If we modified Figure 1(a) by inserting a mediator M or fork F between C and X or between C and Y (but not a fork between both C and X and C and Y), C would remain sufficient (as would M alone, F alone, or any combination of C, M, or F) and the value of expressions (1) and (2) would not change.

In Figure 1(b), the path between X and Y via C ($X \rightarrow C \rightarrow Y$) is direct through C, and so is an open path; hence X and Y may be associated via this path. In both Figures 1(a) and 1(b), the open path will be blocked by conditioning fully on C; hence if X had no direct effect on Y, X and Y would be separated given C, reflecting correctly this absence of effect. Thus, to estimate a total effect of X on Y, we should not condition on C because C is a mediator (intermediate) between X and Y, carrying part of the total effect; but if we want to estimate a direct effect of X on Y, we would condition fully on C. We caution that in applications, further conditioning on confounders of net and direct effects would be needed (Robins & Greenland, 1992; Cole & Hernán, 2002).

Behavior opposite of the confounding case in Figure 1(a) arises in Figure 1(c), where the path between X and Y via C ($X \rightarrow C \leftarrow Y$) is blocked at C, and hence X and Y cannot be associated via this path. This closed path will however be unblocked (opened) by conditioning on C; this means that X and Y may be dependent given C (dependent within at least one level of C) even if there is no effect of X on Y. This would continue to be so if we inserted a mediator M or fork F between X and C or between Y and C (but not a fork between both). Thus, to estimate an effect of X on Y, we should *not* condition on C, the opposite situation from Figure 1(a). Figure 1(c) could arise when C is a complete-record indicator in database studies, an indicator of selection in case-control studies, or a censoring indicator in cohort studies and trials with losses; for a more general graphical treatment of the effect of conditioning on outcome (Y) effects, see Didelez *et al.* (2010) and Bareinboim & Pearl (2011).

7.2 Effects of Conditioning on a Descendant of an Interceptor

Consider next conditioning on a child Z of C for which the connection $C \rightarrow Z$ is not invertible, so that at most conditioning on Z corresponds to only partial adjustment for C. The situations in Figure 1 would arise if (for example) C was unmeasured and Z was an imperfect but non-differential measurement of C or proxy for C (i.e., Z is independent of X and Y given C). Again, if we do not condition on Z in Figures 1(a) and 1(b), X and Y remain connected through C, which is a source of bias for estimating any effect of X on Y in Figure 1(a) or a direct effect in Figure 1(b); in Figure 1(c), X and Y remain separated and there is no bias for estimating any effect of X on Y.

What if we condition on Z only? For each case in Figure 1 we see that Z is connected to X both unconditionally and conditional on Y, and is connected to Y given X. Thus we expect non-collapsibility over Z for all measures. In Figures 1(a) and 1(b), one way to interpret the changes in risk differences and risk ratios is that conditioning on Z partially closed the open path connecting X and Y through C. In Figure 1(c), however, these changes correspond to a partial opening of the unconditionally closed path $X \rightarrow C \leftarrow Y$. See Pearl (2009, p. 338) for a graphical explanation using the concept of "virtual colliders."

Conditioning on Z can be viewed as adjustment for C using a non-differentially misclassified or coarsened (e.g., dichotomized) proxy. Under Figure 1(a), a traditional rule is that such

non-differential proxy adjustment induces partial control of confounding (e.g., Greenland, 1980; Fung & Howe, 1984; Savitz & Baron, 1989); that is, we would ordinarily expect conditioning on Z to move us part way from the confounded unconditional (unadjusted) association of X and Y toward the total effect of X on Y. Nonetheless, Brenner (1993) showed that this rule did not in general extend to polytomous covariates. Ogburn & VanderWeele (2012) showed that exceptions to this rule can occur even for binary covariates, but that the rule holds under mild conditions for binary covariates when using $\text{SMR} = p(y|x_r)/p\{y|x; p(c|x_r)\}$, and also holds for more general adjusted estimators and covariates under monotonicity of the C effects on X and Y ; see also Chaudhuri & Richardson (2003, lemma 3.1). When Z represents a coarsening of C, various algebraic results are available to determine whether conditioning on Z is sufficient for control of confounding by C (Davis, 1986; Ducharme & Lepage, 1986; Geng *et al.* 2001, 2002; Guo *et al.* 2001). Pearl (2010c) provides further conditions under which the X effect on Y is identifiable from Z without C.

By conditioning on Z in Figure 1(b), it appears we are partially adjusting for the effect of X on Y mediated through C; thus, for risk differences and risk ratios, we might ordinarily expect this adjustment to move us partway from the total effect of X on Y toward a direct effect of X and Y. Such movement introduces bias if our target is the total effect, but reduces bias if our target is the direct effect. For a graphical explanation using the concept of "virtual colliders" see Pearl (2009, p. 339).

Again, Figure 1(c) brings an opposite phenomenon from Figure 1(a): Conditioning on Z produces an open non-causal path from X to Y, which we expect to introduce bias where none existed before. In each figure, however, Z is separated from Y by C and X, implying that we will have complete collapsibility over Z given C.

7.3 Effects of Conditioning on an Ancestor of an Interceptor

Turning now to a parent A of C, suppose that C is not completely determined by A (i.e., the connection $A \rightarrow C$ is not perfect), so that at best conditioning on A corresponds to only partial control of C. Unlike a descendant of C, A actually does control C in a causal sense. Hence, under the diagrams we present, the consequences of conditioning on A parallel the consequences that would follow if A were an (imperfect) intervention to set the level of C. Because of this parallel, we will see some telling divergences from what happens when conditioning on the child Z of C: This is because, unlike with A or C, intervention to change the level of Z could have no effect on any other variable.

What if we condition on A only? In Figure 1(a) we see that A (like Z) is connected to X both unconditionally and conditional on Y, and is connected to Y conditional on X, so we expect non-collapsibility for all measures. For risk differences and risk ratios we can interpret these changes as reflecting partial closure of the open path connecting X and Y through C. Thus, for binary A we would ordinarily expect partial control of confounding if we condition on A in place of C, moving us from the confounded unconditional (unadjusted) association of X and Y toward the total effect of X on Y.

In Figure 1(b), however, A is separated from X unconditionally. Thus, from (b), population-standardized measures will be collapsible over A. Nonetheless, A is connected to Y conditional on X, so we expect some A-specific measures to differ from their corresponding unconditional measure, which we might interpret as partial control of the effect of X on Y mediated through C. Furthermore, A is connected to X conditional on Y, so we expect average odds ratios to differ from unconditional odds ratios. It might be tempting to think that these odds-ratio changes represent partial control of the effect of X on Y mediated through C, but the reality is more complex.

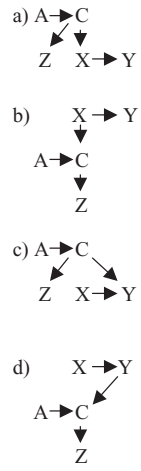

Figure 2. *Graphs with C separated from X or Y under some condition.*

In Figure 1(c), A is not connected to Y conditional on X, and so, unlike with C or Z, we have complete collapsibility over A. In graphical terms, conditioning on A does not even partially open the path from X to Y through C, and thus induces no bias; this is so even if A determines C completely ($C = A$), for in that case X and Y will no longer affect C (so the arrows into C from X and Y disappear). Nonetheless, A is connected to X given C, to X given Y and C, and to Y given X and C. Thus, unlike with Z, we expect non-collapsibility over A given C for all measures. If C represents selection, this means that A will appear to be a confounder among the selected, even though it is not. In contrast, in Figures 1(a) and 1(b), A is separated from Y by C given X, so there will be complete collapsibility over A given C.

The results just described extend to any ancestor A of C that is not connected to X or Y except through A.

7.4 Some Special Cases

We now consider some important special cases that often arise in observational studies of effects. First, suppose C and Y share no ancestor and do not affect each other, except possibly through X. Simple cases are shown in Figures 2(a) and 2(b), which drop the arrows between C and Y in Figures 1(a) and 1(b). A, C, and Z are all separated from Y given X; hence, we have complete collapsibility over A, C, and Z, whether we consider them singly, in pairs, or all together. Figure 2(a) is of special note because A, C, and Z could singly appear important in a propensity-score analysis of the effect of X on Y, yet none would be needed or even desirable for any adjustment procedure; see the discussion later on bias amplification and variance inflation.

Next, suppose C and X share no ancestor and do not affect each other, except possibly through Y. Simple cases are shown in Figures 2(c) and 2(d), which drop the arrows between C and X in Figures 1(a) and 1(c). In Figure 2(c), A, C, and Z are all unconditionally separated from X; hence population-standardized measures of dependence of Y on X are collapsible over A, C, or Z, since $p\{y|x; p(c)\} = p\{y|x; p(a)\} = p\{y|x; p(z)\} = p(y|x)$. Moreover, those measures will be unconfounded, since $p(y|x) = p(y|\text{do}[x])$. Thus, A, C, and Z could singly appear important in a Y-regression analysis of the effect of X on Y, yet none would be needed or even desirable for any adjustment procedure. Nonetheless, C, A, and Z are all connected to X given Y and to Y

given X; hence we should expect the XY odds ratios to be non-collapsible over them despite the lack of confounding (illustrating graphically the "odds-ratio paradox" discussed earlier).

In Figure 2(d), C and Z are connected to X unconditionally and to Y given X; hence we expect population-standardized measures of dependence of Y on X to be non-collapsible over C and Z, and this remains so if we condition on A as well. On the other hand, A, C, and Z are separated from X given Y, implying that odds ratios will be collapsible over all of them. Furthermore, if we do not condition on C or Z, A will be separated from Y conditional on X and so we have complete collapsibility over A.

Figure 2(d) can be taken as representing a case-control study with exposure X, disease Y, C indicating selection, and Z indicating consent. In such a study, Y by definition affects selection C very strongly, resulting in severe non-collapsibility over C of all measures except odd ratios. Under this graph the unconditional measures are not confounded, hence this non-collapsibility over C represents a strong bias from conditioning on C. This bias afflicts all familiar measures that depend on absolute frequencies of Y values in some fashion, such as risk differences, odds differences, and risk ratios; for example, risk ratios cannot exceed 2 if the absolute frequency of $Y = 1$ is never below $\frac{1}{2}$, as often occurs in case-control studies. In contrast, odds ratios relating X to Y depend only on relative frequencies of Y values given X and C, and hence are collapsible, as can also be seen from the fact that C and Z are separated from X by Y ; this collapsibility can be viewed as a graphical generalization of the famous result by Cornfield (1951), and justifies use of the odds ratios from participants ($C = Z = 1$) to estimate the unconditional odds ratios (Didelez *et al.*, 2010; Bareinboim & Pearl, 2011).

Nonetheless, an effect of X (or of an ancestor of X) on selection C (as in Figure 1(c)) or consent Z will connect X to Y, and thus introduce bias in the odds ratio as well as in other measures; this bias is the familiar Berksonian form of selection bias (Greenland *et al.*, 1999a; Glymour & Greenland, 2008; Pearl, 2009). Similar concerns arise in cohort studies in which C represents loss to follow-up or other forms of censoring, in studies in which C indicates completeness of records, and in trials in which C is a compliance indicator and the analysis discards non-compliers ("per-protocol" analysis).

8 Extensions to Diagrams with Other Adjustment Variables

There are many ways to extend the previous graphical results. We present some examples to illustrate how the rules we have described may guide us in selecting adjustment variables that are not confounders in the classical sense that C is in Figure 1(a). In each example in Figures 3(a)–3(d), the XY association exhibits some form of non-collapsibility over C, but in the first example this non-collapsibility amplifies a bias, in the second and third it reduces a bias, and in the fourth it does both. We again emphasize that biases within these simple structures extend to graphs that contain these structures as subgraphs.

8.1 Bias Amplification

Figure 3(a) adds an uncontrolled confounder U to Figure 2(a). A, C, and Z satisfy graphical conditions for instrumentality, that is, unconditional connection to X and connection to Y only through X as a mediator (Pearl, 2009); they are also connected to X given Y. Unlike in Figure 2(a), however, A, C, and Z are connected to Y conditional on X via the path $C \rightarrow X \leftarrow U \rightarrow Y$ because X is a collider on that path. Thus, we expect non-collapsibility over A, C, and Z, singly and jointly, for all measures (if we could condition on U we would be back in a completely collapsible situation like that in Figure 2(a)).

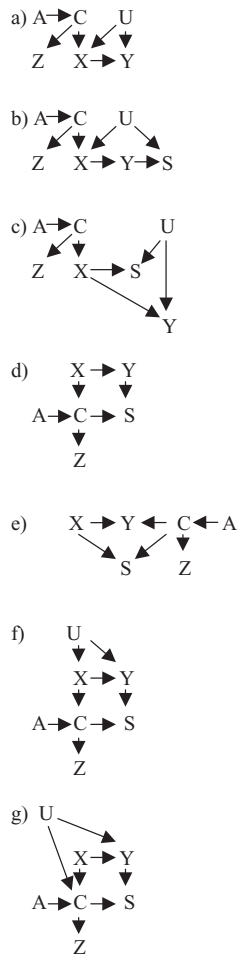

Figure 3. *Graphs with an additional ancestor U or descendant S of X or Y.*

Considering cases in which effects can be given a sign (positive or negative), Bhattacharya & Vogt (2007) and Pearl (2010d) show algebraically how the unconditional non-collapsibility over A, C, Z in Figure 3(a) can lead to increased bias from the confounding back-door path $X \leftarrow U \rightarrow Y$. In particular, in the class of models analyzed in Pearl (2010d), the bias from adjusting for A, C, or Z is in the same direction as (and thus amplifies) the bias from failing to adjust for U. Intuitively, when we consider the unconditional (crude) association between X and Y, systematic variation in X is partly explained by variation in C and partly by variation in U. The U component is transmitted to Y via the confounding path $X \leftarrow U \rightarrow Y$ and so counts toward bias. If we condition on C, however, the C component vanishes; hence all systematic variation in X comes from variation in U and is transmitted to Y via the confounding path, with larger bias as a result. Such amplification may also take place when C is a "near instrument," that is, a confounder that influences exposure much more strongly than the outcome (Pearl, 2011).

A test of collapsibility over an instrumental variable (C, A, or Z in Figure 3(a)) provides a test of sufficiency of those covariates already controlled. Given insufficiency and due attention to instrumentality (e.g., Hernán & Robins, 2006), we may turn to specialized methods

for instrumental-variable adjustments (e.g., Sommer & Zeger, 1991). The bias-amplification problem does not arise in these methods because the instruments are used to correct the unconditional association, instead of being conditioned on as in outcome-regression and propensity-score adjustment. Nonetheless, some authors recommend selecting all variables that influence exposure X for propensity-score adjustments, without regard to their relation to the outcome Y (e.g., Hirano & Imbens, 2001; Rubin, 2002, 2009). Unfortunately adjusting for variables related only to exposure may not only amplify bias, but may unnecessarily inflate variances (e.g., see Day *et al.*, 1980; Thomas & Greenland, 1983; Brookhart *et al.*, 2006; Austin *et al.*, 2007; White & Lu, 2011; de Luna *et al.*, 2011).

Conditioning on apparent instrumental variables can also amplify certain types of selection bias (Pearl, 2010d). Consider Figure 3(b), which modifies Figure 3(a) by replacing $U \rightarrow Y$ with $U \rightarrow S \leftarrow Y$. Before conditioning on S, the association of X and Y will be collapsible over A, C, or Z because X separates those variables from Y. Nonetheless, conditioning on S opens a path from X to Y via U and S, introducing selection bias. Furthermore, X no longer separates A, C, or Z from Y, so we should expect the association of X and Y to be non-collapsible over A, C, or Z given S. This non-collapsibility over selection S may result in either amplification or attenuation of bias.

On the other hand, if A, C, and Z remain separated from Y by X after selection, conditioning on A, C, or Z will not amplify the selection bias. This is because A, C, and Z will remain independent of Y given X. As an example, consider Figure 3(c), which modifies Figure 3(a) by replacing $U \rightarrow X$ with $U \rightarrow S \leftarrow X$. Again, conditioning on S produces bias because it opens a path from X to Y via U and S, but the association remains unchanged and hence is not further biased by conditioning on A, C, or Z. This example illustrates bias equivalence (bias from conditioning on S is equivalent to bias from conditioning on S and any combination of A, C, or Z), which is discussed below.

8.2 Selection-Bias Removal

Figure 3(d) modifies Figure 1(c) by replacing $X \rightarrow C \leftarrow Y$ with $X \rightarrow C \rightarrow S \leftarrow Y$, as could arise when S is an indicator of inclusion in the statistical analysis ($S = 1$ for inclusion) and C is an adherence or complete-record indicator. The unconditional XY association is unbiased for the effect of X on Y. S is connected to X both unconditionally and given Y, and is connected to Y given X, so we expect conditioning on S alone (as in per-protocol and complete-case analyses) to introduce bias in estimating the effect (Pearl, 2009, Ch.8; Daniel *et al.*, 2011; Westreich, 2011).

As in Figure 1(c), in Figure 3(d) both C and Z are connected to X unconditionally and given Y, while A is not connected to X. Nonetheless, A, C, and Z are independent of Y given X, so we have complete collapsibility over them all. Conditioning on S reverses the situation, however: A, C, and Z are connected to Y given X and S, and are connected to X given S and given Y and S. Thus, we expect non-collapsibility over A, C, and Z given S, singly and jointly. For the odds ratio, however, further conditioning on C removes this problem: S is separated from X given C and Y, so we have collapsibility of the XY odds ratio over S given C as well as over C; hence this odds ratio is collapsible over the compound variable $\{C,S\}$ even though we cannot assume that it is collapsible over S or over C given S. Put another way, conditioning on C removes the selection bias in the odds ratio produced by conditioning on S, making C a "bias-breaking" variable for the odds ratio (Geneletti *et al.*, 2009; Didelez *et al.*, 2010; Bareinboim & Pearl, 2011).

The same bias removal occurs if the $X \rightarrow C$ relation is reversed to $C \rightarrow X$ so that C is a fork rather than a mediator between X and S ($X \leftarrow C \rightarrow S$), or if C is instead a mediator between Y and S ($Y \rightarrow C \rightarrow S$) (Bareinboim & Pearl, 2011). Note, however, that in all these cases, removal

of bias by conditioning on C is limited to the odds ratio; that is, we expect only odds-ratio collapsibility over $\{C,S\}$. This limitation corresponds to the well-known fact that conditioning on a variable affected by the outcome variable Y (as in case-control sampling) will alter the observed proportions with a specific outcome (such as disease) and so alter risk differences, risk ratios, and odds differences (Rothman *et al.*, 2008, ch. 8; Pearl 2009, p. 339).

In Figure 3(d), we expect non-collapsibility of the XY odds ratio over A and over Z given S, but (in contrast to $\{C,S\}$) we also expect non-collapsibility over $\{A,S\}$, $\{Z,S\}$, and $\{A,Z,S\}$. This means that, after conditioning on S, we might ordinarily expect bias reduction from the change induced by conditioning on A, Z, or both, but we would not expect complete bias removal.

Consider next Figure 3(e), where C is instead a fork between Y and S ($Y \leftarrow C \rightarrow S$). Again there is no bias in any unconditional measure, but we expect all measures to be non-collapsible over S and thus biased conditional on S. Further conditioning on C will remove this bias from all the C-specific measures (this time, not just the odds ratios); that is, we expect C-specific measures to be collapsible over S. Furthermore, we might ordinarily expect conditioning on A, Z, or both to partially remove the bias from conditioning on S. Typically, however, one observes only the distribution of C among the selected, $p(c|S = 1)$; thus the population-standardized measures cannot be recovered without further information to reconstruct $p(c)$ (Bareinboim & Pearl, 2011).

Finally, in all the cases just described (Figure 3(d) and modifications, and 3(e)), one could reconstruct all desired measures from selection-biased data if given selection probabilities $p(S = 1|R)$ where R is a set of variables separating S from X and Y (R may include X or Y or both). These probabilities allow reconstruction of the (unbiased) marginal distribution $p(x, y)$, for example, by using their inverses to reweight the data or to construct direct bias corrections (Horvitz & Thompson, 1952; Rothman *et al.*, 2008, pp. 362–363).

8.3 Bias Equivalence

Figure 3(f) adds an uncontrolled unconditional confounder U of XY to Figure 3(d). Now the unconditional XY odds ratio is biased, being a mix of the study effect $X \rightarrow Y$ and the association over the confounding back-door path $X \leftarrow U \rightarrow Y$. Because this confounding path has no overlap with the selection-bias path $X \rightarrow C \rightarrow S \leftarrow Y$, the previous observations about the latter path continue to apply: We have non-collapsibility over S but collapsibility over both C and $\{C,S\}$ relative to the U-confounded (unconditional) XY odds ratio. This collapsibility is a simple example of confounding equivalence (Pearl & Paz, 2010): we are left with the same degree of confounding (from U) whether we condition on nothing, on C, or on both C and S.

More generally, Pearl & Paz (2010) derive conditions for determining when two arbitrary covariate sets R and S are confounding equivalent in the sense that standardization by either $p(r)$ or $p(s)$ yields the same association or effect measure. Because of the correspondence between probability standardization and inverse-probability data weighting described earlier, such standardization equivalence will also imply equivalence of marginal effects derived from weights of $1/p(x|r)$ and weights of $1/p(x|s)$. Remarkably, this equivalence can occur even if the associations are heterogeneous across strata of R and S; but in that case we should not expect perfect equality between R-adjusted and S-adjusted measures that correspond to other weighting schemes (e.g., logistic-regression coefficients).

Conditional on U we also have non-collapsibility over S but collapsibility over both C and $\{C, S\}$ relative to the unconfounded (U-conditional) XY odds ratio, so we also have bias equivalence given U (which in this case is no bias whether in addition to U we condition on nothing, on C, or on both C and S). If instead we condition on A, Z or both after conditioning on S, we no longer have such equivalencies, since we have unconditional collapsibility over A,

Z, or both, but we expect non-collapsibility over $\{A, S\}$, $\{Z, S\}$, and $\{A, Z, S\}$. Thus as in Figures 3(d) and 3(e), we would not expect conditioning on A or Z to be sufficient for removal of the bias from conditioning on S, even for odds ratios.

8.4 Overlapping Bias Paths

Figure 3(g) adds an uncontrolled variable U to Figure 3(d), one which does not unconditionally confound the XY relation but does confound other relations. Hence, there is no bias unconditionally. Nonetheless, conditioning on C, S, or Z now opens a new path from X to Y, $X \rightarrow C \leftarrow U \rightarrow Y$. As a consequence, we no longer have collapsibility over C or Z, and conditioning on S opens two paths from X to Y (the new path, as well as $X \rightarrow C \rightarrow S \leftarrow Y$).

As in Figure 3(f), we still have odds-ratio collapsibility over S given C, so C and $\{C, S\}$ remain bias equivalent for odds ratios, as do $\{C, U\}$ and $\{U, C, S\}$; and, once we condition on S (as we are forced to do when S is selection), we would have to condition on U as well as C to remove all odds-ratio bias. Unlike Figure 3(f), however, in Figure 3(g) U could be ignored if there was no conditioning on C, S, or Z. Furthermore, we might ordinarily expect the bias from the $X \rightarrow C \leftarrow U \rightarrow Y$ path to be larger if C were conditioned than if only S were conditioned. In this sense, after conditioning on S, we would ordinarily expect further conditioning on C to amplify the bias from the $X \rightarrow C \leftarrow U \rightarrow Y$ path even though it would remove the bias from the $X \rightarrow C \rightarrow S \leftarrow Y$ path; the net impact of conditioning on C given S is thus hard to predict.

If instead of C we condition on A, Z or both after conditioning on S, we no longer have bias equivalencies. We have unconditional collapsibility over A, but after conditioning on S we expect non-collapsibility over any combination from A, U, or Z. Thus, as in Figures 3(d)–(f), we would not expect conditioning on A, Z or both to be sufficient for complete removal of the bias from conditioning on S, even after conditioning on U.

As seen in Figures 3(f) and 3(g), odds ratios have the potential to remain unbiased when conditioning on variables affected by the outcome Y, provided that conditioning does not open a path from X to Y. The application of these results extends from odds-ratio to rate-ratio analysis when sampling or conditioning is done in a manner that forces sample odds ratios to estimate hazard (rate) ratios, as is typical in case-control studies with risk-set (density) sampling and in survival analysis (Rothman *et al.*, 2008, pp. 113–114 and 294–295).

9 Separation and Collapsibility Testing

The fact that different graphical and probability structures have different collapsibility implications leads naturally to consideration of model testing and selection. A causal model can only be tested through its statistical implications, such as the conditional independencies implied by separation; as a consequence, the bulk of causal assumptions embedded in such a model will remain untested. Nonetheless, many testing strategies have been developed (Pearl, 1988, ch. 8; Robins, 1999; Spirtes *et al.*, 2001, sec. 6.9; Tian & Pearl, 2002; Shpitser & Pearl, 2008; Pearl, 2009, pp. 345–348; de Luna *et al.*, 2011; Shpitser *et al.*, 2011).

To test a separation criterion, higher statistical power can be attained by testing the independency implied by the criterion rather than by testing the implied collapsibility. On the other hand, collapsibility is often more relevant to causal inference (as may be seen from the examples later). Thus, Pearl & Paz (2010) suggest using collapsibility tests and bias-equivalence tests as diagnostics for graphical models, analogous to collapsibility-based tests of parametric models (e.g., Whittemore, 1978; Hausman, 1978; Greenland & Mickey, 1988; Clogg *et al.*,

1995). Their recommendation is based on the fact that collapsibility holds under either or both of two conditional independencies (as well as under other conditions); therefore, if a test rejects collapsibility, it rejects all graphical models that imply the collapsibility.

When C and S are vectors of covariates, such tests can be performed using familiar modeling strategies. Suppose a graph predicts that a measure of the dependence of the outcome Y on the exposure X is collapsible over C given S. One approach starts with a model for $p(y|x, c, s)$, such as a logistic regression model, and then tests whether the X coefficient is equal to that obtained when C is dropped to produce a model for $p(y|x, s)$; equality is collapsibility of the X coefficient over C (Clogg *et al.*, 1995).

Asymptotic tests can however falter with very high-dimensional S, especially when S strongly predicts X and Y. An alternative for these cases replaces the pair of vectors (C, S) and the vector S with fitted values from models for the exposure-propensity scores $p(x|c, s)$ and $p(x|s)$, respectively; that is, (C, S) and S are replaced by their fitted X-propensity scores. We may then test equality of the adjusted measures derived from the two scores, which is equivalent to testing collapsibility of the measure over C (Pearl, 2009, p. 349). The two approaches can be combined by using the propensity scores to fit the Y (outcome) model, as in doubly robust estimation (Kang & Shafer, 2007). Again, rejection of equality (collapsibility) of a measure after deleting C from both the Y and X models implies rejection of all graphical models that entail collapsibility of the X coefficient over C (assuming correct model specification).

We caution that the use of preliminary tests for model and covariate selection (whether for independence or for collapsibility testing) can distort the final p-values and confidence intervals for the effect of interest. Among commonly proposed solutions are methods that rely on shrinkage instead of selection; for example, see Greenland (2008) for a review and suggested alternatives to preliminary testing in the context of confounder selection in regression modeling of effects. For high-dimensional problems, however, strategies based on cross-validation and simulation, such as bootstrapping, bagging, and boosting (Hastie *et al.*, 2009) exhibit considerable advantages. To date, their application to causal modeling has been mostly for inverse-probability weight estimation (e.g., McCaffrey *et al.*, 2004; Lee *et al.*, 2011) although more targeted uses have been developed (Van der Laan & Rose, 2011).

10 Discussion

We have reviewed algebraic results and introduced graphical criteria to answer questions about when one may expect adjustment for particular variables to alter the bias from a particular source in a given graphical model. In connection with the model-selection issue, there is considerably more that could be researched and discussed regarding implications of adjustment for statistical efficiency and mean-squared error (or more generally, net loss), and quantification of the bias added or removed by a given adjustment. Basic results on these topics are available, especially for ratio measures (e.g., Yanagawa, 1984; Gail, 1986; Robinson & Jewell, 1991; Greenland, 1991, 2003; Clogg *et al.*, 1995; De Stavola & Cox, 2008; Janes *et al.*, 2010; de Luna *et al.*, 2011), but many details and extensions remain to be worked out (which is unsurprising given the many parameters that must be modeled to quantify efficiency and bias). A systematic review of this topic would be valuable for guidance on currently available methods as well as for future research.

It has been observed that, under simple models in which each association can be fully summarized by a single number, association and hence non-collapsibility appears to attenuate when it arises from more extended paths (Greenland, 2003; Chaudhuri & Richardson, 2003). Under these conditions (which typify models in the health and social sciences), we would

ordinarily expect the strength of associations of Z with X and Y to be less than the strength of associations of C with X and Y. That is because each path connecting Z to X or Z to Y properly contains the corresponding path connecting C to X or C to Y. As a result, we would ordinarily expect a smaller degree of XY non-collapsibility over Z than over C, which means that adjusting for Z will move us less from the unconditional association than will adjusting for C. In Figure 1(a), this means we would ordinarily not expect Z-adjustment to remove as much bias as C-adjustment; in Figure 1(c), it means we would ordinarily not expect Z adjustment to produce as much bias as C-adjustment; and in Figure 1(b) the bias implication depends on whether we are interested in a direct or total effect.

Another avenue for extending qualitative results is in terms of direction of bias, which can be derived by adding signs to path arrows (VanderWeele *et al.*, 2008; VanderWeele & Robins, 2010). Quantitative considerations will have to enter when one considers multiple bias sources, as occur in Figures 3(f) and 3(g) after conditioning on S. We expect that the net bias in most such situations will be not be simple in form and will be heavily dependent on contextual details; thus general results that can simplify context-specific analyses would be valuable. We hope that the results provided here provide a reasonable starting or reference point for further extensions.

Acknowledgements

We wish to thank Charles Poole for prompting this investigation with penetrating questions, Tyler VanderWeele for helpful comments and correspondence on the topic, and Thomas Richardson and the referees for unusually detailed comments that greatly improved the presentation.

References

Austin, P.C., Grootendorst, P. & Anderson, G.M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Stat. Med.*, **26**, 734–753.

Bareinboim, E. & Pearl, J. (2011). Controlling Selection Bias in Causal Inference. UCLA Cognitive Systems Laboratory, Technical Report R-381, available at http://ftp.cs.ucla.edu/pub/stat_ser/r381.pdf.

Bareinboim, E., Brito, C. & Pearl, J. (2011). Local characterizations of causal Bayesian networks. In *Proceedings of the Second International IJCAI Workshop on Graph Structures for Knowledge Representation and Reasoning*, vol. 1, Eds. M. Croitoru, O. Corby, J. Howse, S. Rudolph & N. Wilson, pp. 6–11. Barcelona: GKR.

Bhattacharya, J. & Vogt, W. (2007). Do Instrumental Variables Belong in Propensity Scores? Technical Working Paper 343, National Bureau of Economic Research, Cambridge, MA, available at http://www.nber.org/papers/t0343.

Brenner, H. (1993). Bias due to non-differential misclassification of polytomous confounders. *J. Clin. Epidemiol.*, **46**, 57–63.

Brookhart, M., Schneeweiss, S., Rothman, K.J., Glynn, R., Avorn, J. & Stürmer, T. (2006). Variable selection for propensity score models. *Amer. J. Epidemiol.*, **163**, 1149–1156.

Chaudhuri, S., Richardson, T.S. (2003). Using the structure of d-connecting paths as a qualitative measure of the strength of dependence. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, pp. 116–123. Corvallis, WA: AUAI.

Clogg, C.C., Petkova, E. & Haritou A. (1995). Statistical methods for comparing regression coefficients between models (with discussion). *Amer. J. Sociol.*, **100**, 1261–312.

Cole, S.R. & Frangakis, C.E. (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, **20**, 3–5.

Cole, S. & Hernán, M. (2002). Fallibility in estimating direct effects. *Int. J. Epidemiol.*, **31**, 163–165.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data: application to cancer of the lung, breast and cervix. *J. Nat. Cancer Inst.*, **11**, 1269–1275.

Daniel, R.M., Kenward, M.G., Cousens, S.N. & De Stavola, B.L. (2011). Using causal diagrams to guide analysis in missing data problems. *Statist. Meth. Med. Res.*, **20**, in press.

Davis, L.J. (1986). Whittemore's notion of collapsibility in multidimensional contingency tables. *Comm. Statist. Theory Methods*, **15**, 2541–2554.

Day, N.E., Byar, D.P. & Green, S.B. (1980). Overadjustment in case-control studies. *Amer. J. Epidemiol.*, **112**, 696–706.
de Luna, X., Thomas, S., Richardson, T.S. & Waernbaum, I. (2011). Covariate selection for the non-parametric estimation of an average treatment effect. *Biometrika*, in press.
De Stavola, B.L. & Cox, D.R. (2008). On the consequences of overstratification. *Biometrika*, **95**, 992–996.
Didelez, V., Kreiner, S. & Keiding, N. (2010). On the use of graphical models for inference under outcome dependent sampling. *Statist. Sci.*, **25**, 368–387.
Ducharme, G.R. & Lepage, Y. (1986). Testing collapsibility in contingency tables. *J. R. Stat. Soc. Ser. B*, **48**, 197–205.
Fung, K.Y. & Howe, G.R. (1984). The effect of joint misclassification of risk factors and confounding factors upon estimation and power. *Int. J. Epidemiol.*, **13**, 366–370.
Frydenberg, M. (1990). Marginalization and collapsibility in graphical statistical models. *Ann. Statist.*, **18**, 790–805.
Gail, M.H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In *Modern Statistical Methods in Chronic Disease Epidemiology*, Eds. S. H. Moolgavkar & R.L. Prentice, pp. 3–18. New York: Wiley.
Geneletti, S., Ricequalityson, S. & Best, N. (2009). Adjusting for selection bias in retrospective case-control studies. *Biostatistics*, **10**, 17–31.
Geng, Z. (1992). Collapsibility of relative risk in contingency tables with a response variable. *J. R. Stat. Soc.Ser. B*, **54**, 585–593.
Geng, Z. & Li, G. (2002). Conditions for non-confounding and collapsibility without knowledge of completely constructed causal diagrams. *Scand. J. Stat.*, **29**, 169–181.
Geng, Z., Guo, J.H., Lau, T.S. & Fung, W.K. (2001). Confounding, homogeneity and collapsibility for causal effects in epidemiologic studies. *Statist. Sinica*, 11, 63–75.
Geng, Z., Guo, J.H & Fung, W.K. (2002). Criteria for confounders in epidemiological studies. *J. R. Stat. Soc.Ser. B*, **64**, 3–15.
Glymour, M.M. & S. Greenland (2008). Causal Diagrams. Ch. 12. In *Modern Epidemiology*, Eds. K.J. Rothman, S. Greenland & T.L. Lash, 3rd ed. Philadelphia: Lippincott.
Greenland, S. (1980). The effect of misclassification in the presence of covariates. *Amer. J. Epidemiol.*, **112**, 564–569.
Greenland, S. (1991). Reducing mean squared error in the analysis of stratified epidemiologic studies. *Biometrics*, **47**, 773–775.
Greenland, S. (1996). Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*, **7**, 498–501.
Greenland, S. (2003). Quantifying biases in causal models: classical confounding versus collider-stratification bias. *Epidemiology*, **14**, 300–306.
Greenland, S. (2005a). Epidemiologic measures and policy formulation: lessons from potential outcomes (with discussion). *Emerging Themes in Epidemiology* (online journal) 2:1–4. (Originally published as "Causality theory for policy uses of epidemiologic measures," Chapter 6.2. In *Summary Measures of Population Health*, Eds. C.J.L. Murray, J.A. Salomon, C.D. Mathers & A.D. Lopez (2002), pp. 291—302. Cambridge, MA: Harvard University Press/WHO.).
Greenland, S. & Mickey, R.M. (1988). Closed-form and dually consistent methods for inference on collapsibility in 2xJxK tables. *Appl. Stat.*, **37**, 335–343.
Greenland, S. & Pearl, J. (2010). Causal diagrams. In *International Encyclopedia of Statistical Sciences*, Ed. M. Lovric. New York: Springer.
Greenland, S., Pearl, J. & Robins, J.M. (1999a). Causal diagrams for epidemiologic research. *Epidemiology*, **10**, 37–48.
Greenland, S., Robins, J.M. & Pearl, J. (1999b). Confounding and collapsibility in causal inference. *Statist. Sci.*, **14**, 29–46.
Greenland, S. (2010). Overthrowing the tyranny of null hypotheses hidden in causal diagrams. Ch. 22. In *Heuristics, Probabilities, and Causality: A Tribute to Judea Pearl*, Eds. R. Dechter, H. Geffner & J.Y. Halpern, pp. 365–382. London: College Press.
Guo, J.H. & Geng, Z. (1995). Collapsibility of logistic regression coefficients. *J. R. Stat. Soc. Ser. B*, **57**, 263–267.
Guo, J., Geng, Z. & Fung, W.-K. (2001). Consecutive collapsibility of odds ratios over an ordinal background variable. *J. Multivariate Anal.*, **79**, 89–98.
Halloran, M.A. & Struchiner, C.J. (1995). Causal inference for infectious diseases. *Epidemiology*, **6**, 142–151.
Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer.
Hausman, J.A. (1978). Specification tests in econometrics. *Econometrica*, **43**, 1251–1271.
Hernán, M.A. (2005). Hypothetical interventions to define causal effects—afterthought or prerequisite? *Amer. J. Epidemiol.*, **162**, 618–620.
Hernán, M.A. & Robins, J.M. (2006). Instruments for causal inference. *Epidemiology*, **17**, 360–372.

Hernán M.A., Hernandez-Diaz, S. & Robins, J.M. (2004). A structural approach to selection bias. *Epidemiology*, **15**, 615–625.

Hirano, K. & Imbens, G.W. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv. Outcomes Res. Methodol.*, **2**, 259–278.

Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Stat. Assoc.*, **47**, 663–685.

Janes, H., Dominici, F. & Zeger, S. (2010). On quantifying the magnitude of confounding. *Biostatistics*, **11**, 572–582.

Joffe, M.M., Yang, W.P. & Feldman, H.I. (2010). Selective ignorability assumptions in causal inference. *Int. J. Biostat.*, **6**(2), article 11, available at http://www.bepress.com/ijb/vol6/iss2/11.

Kang, J.D., Shafer, J.L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statist. Sci.*, **22**, 523–580.

Kiiveri, H., Speed, T.P., Carlin, J.B. (1984) Recursive causal models. *J. Austral. Math. Soc. Ser. A*, **36**, pp. 30–52.

Lauritzen, S. (1996). *Graphical Models*. Oxford: Clarendon Press.

Lauritzen, S.L. & D.J. Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. R. Stat. Soc. Ser. B*, **50**, 157–224.

Lee, B.K., Lessler, J. & Stuart E.A. (2011) Weight trimming and propensity score weighting. *PLoS One*, **6**(3), 1–6: e18174. doi:10.1371/journal.pone.0018174.

McCaffrey, D.F., Ridgeway, G. & Morral, A.R. (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Meth.*, **9**, 403–425.

Miettinen, O.S. & Cook E.F. (1981). Confounding: essence and detection. *Amer. J. Epidemiol.*, **114**, 593–603.

Ogburn, E.L. & VanderWeele, T.J. (2012). On the nondifferential misclassification of a binary confounder. *Epidemiology*, in press.

Pearl, J. (1986). Fusion, propagation and structuring in belief networks. *Artif. Intell.*, **9**, 241–288.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika*, **82**, 669–710.

Pearl, J. (2009). *Causality*, 2nd ed. New York: Cambridge University Press.

Pearl, J. (2010a). An introduction to causal inference. *Int. J. Biostat.* (online journal), **6**(2), Article 7, available at http://www.bepress.com/ijb/vol6/iss2/7.

Pearl, J. (2010b). On the consistency rule in causal inference. *Epidemiology*, **21**, 872–875.

Pearl, J. (2010c). On measurement bias in causal inference. In *Proceedings of the 26th Conference on Uncertainty and Artificial Intelligence*, Eds. P. Grunwald & P. Spirtes, pp. 425–432. Corvallis, WA: AUAI Press. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r357.pdf.

Pearl, J. (2010d). On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pp. 417–427. Corvallis, WA: AUAI Press. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r356.pdf.

Pearl, J. (2011). Understanding bias amplification. *Amer. J. Epidemiol.*, available at http://ftp.cs.ucla.edu/pub/stat_ser/r386.pdf.

Pearl, J. & Paz, A. (2010). Confounding equivalence in observational studies. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pp. 433–441. Corvallis, WA: AUAI Press. Available at http://ftp.cs.ucla.edu/pub/stat ser/r343.pdf.

Robins, J.M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling with Applications to Causality*, Ed. M. Berkane, pp. 69–117. New York: Springer-Verlag.

Robins J.M. (1999). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In *Computation, Causation, and Discovery*, Eds. C. Glymour & G. Cooper, pp. 349–405. Menlo Park, CA/Cambridge, MA: AAAI Press/The MIT Press.

Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology*, **12**, 313–320.

Robins, J.M. & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.

Robins, J.M. & Richardson, T.S. (2010). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, Ed. P. Shrout. New York: Oxford University Press.

Robins J.M., Hernán M.A. & Brumback B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.

Robinson, L.D. & Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models. *Int. Statist. Rev.*, **2**, 227–240.

Rothman, K.J., Greenland S. & Lash T.L., eds. (2008). *Modern Epidemiology*, 3rd ed. Philadelphia: Lippincott.

Rubin, D.B. (2002). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.*, **2**, 169–188.

Rubin, D.B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Stat. Med.*, **28**, 1420–1423.
Samuels, M.L. (1981). Matching and design efficiency in epidemiological studies. *Biometrika*, **68**, 577–588.
Sato, T. & Matsuyama, Y. (2003). Marginal structural models as a tool for standardization. *Epidemiology*, **14,** 680–686.
Savitz, D.A. & Baron, A.E. (1989). Estimating and correcting for confounder misclassification. *Amer. J. Epidemiol.*, **129**, 1062–1071.
Schisterman, E., Cole, S. & Platt, R. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, **20**, 488–495.
Sommer, A.S. & Zeger, S. (1991). On estimating efficacy from clinical trials. *Stat. Med.*, **10**, 45–52.
Shpitser, I. & Pearl, J. (2008). Dormant independence. In *Proceedings of the 23rd Conference on Uncertainty and Artificial Intelligence*, pp. 1081–1087. Corvallis, WA: AUAI Press. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r340.pdf.
Shpitser, I. & Pearl, J. (2009). Effects of treatment on the treated: identification and generalization. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Eds. J. Bilmes & A. Ng, pp. 514–521. Corvallis, WA: AUAI Press.
Shpitser, I., VanderWeele, T.J. & Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Conference on Uncertainty and Artificial Intelligence*, Eds. P. Grunwald & P. Spirtes, pp. 527–536. Corvallis, WA: AUAI Press. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r340.pdf.
Shpitser, I., Richardson, T.S. & Robins, J.M. (2011). An efficient algorithm for computing interventional distributions in latent variable causal models. In *Proceedings of the 27th Conference on Uncertainty and Artificial Intelligence*. Corvallis, WA: AUAI Press, in press.
Spiegelhalter, D. (1987). Coherent evidence propagation in expert systems. *J. R. Stat. Soc. Ser. D* (*The Statistician*), **36**, 201–210
Spirtes, P., Glymour, C. & Scheines, R. (2001). *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA: MIT Press.
Thomas, D.C. & Greenland, S. (1983). The relative efficiencies of matched and independent sample designs for case-control studies. *J. Chronic Dis.*, **36**, 685–697.
Tian, J. & Pearl, J. (2002). On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pp. 61–68. Menlo Park, CA: AAAI Press/The MIT Press. Available at http://ftp.cs.ucla.edu/pub/stat_ser/R305.pdf.
Tian, J., Paz, A. & Pearl, J. (1998). Finding a minimal d-separator. UCLA Cognitive Systems Laboratory, Technical Report (R-254), Feb. 1998. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r254.pdf
van der Laan, M.J. & Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data.* New York: Springer.
VanderWeele, T.J. (2009a). On the relative nature of over-adjustment and unnecessary adjustment. *Epidemiology*, **20**, 496–499.
VanderWeele, T.J. (2009b). Concerning the consistency assumption in causal inference. *Epidemiology*, **20**, 880–883.
VanderWeele, T.J. & Robins, J.M. (2010). Signed directed acyclic graphs for causal inference. *J. R. Stat. Soc. Ser. B*, **72**, 111–127.
VanderWeele, T.J., Hernán, M.A. & Robins, J.M. (2008). Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology*, **19**, 720–728.
Wermuth, N. (1987). Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable. *J. R. Stat. Soc. Ser. B*, **49**, 353–364.
Westreich, D. (2011). Berkson's bias, selection bias, and missing data. *Epidemiology*, in press.
White, H. & Lu, X. (2011), Causal diagrams for treatment effect estimation with application to efficient covariate selection. *Rev. Econ. Stat.*, in press.
Whittemore, A.S. (1978). Collapsibility of multidimensional contingency tables. *J. R. Stat. Soc. Ser. B*, **40**, 328–340.
Yanagawa, T. (1984). Case-control studies: assessing the effect of a confounding factor. *Biometrika*, **71**, 191–194.
Yule, G.U. (1934). On some points related to vital statistics, more especially statistics of occupational mortality. *J. R. Stat. Soc.*, **97**, 1–84.

Résumé

Nous passons en revue les méthodes probabilistes et graphiques dans la détection de situations où la dépendance d'une première variable par rapport à une seconde variable se trouve modifiée par la prise en compte d'une troisième

variable, que cette dépendance soit de nature causale ou purement prédictive. Nous mettons l'accent, en particulier, sur la détection des cas où la prise en compte d'une tierce variable entraîne la réduction ou l'augmentation du biais des mesures d'association représentant l'impact causal d'une variable sur l'autre. Nous considérons ensuite les situations dans lesquelles la tierce variable est susceptible d'annuler, ou de biaiser l'estimation des effets de causalité, ainsi que quelques cas particuliers utiles dans les études de cas-témoins, les études de cohortes, et les essais en présence de non-observance (non-compliance).

[*Received November 2010, accepted August 2011*]