WHY THERE IS NO STATISTICAL TEST FOR CONFOUNDING, WHY MANY THINK THERE IS, AND WHY THEY ARE ALMOST RIGHT*

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

1 INTRODUCTION

Confounding is a simple concept. If we undertake to estimate the effect of one variable (X) on another (Y) by examining the statistical association between the two, we ought to ensure that the association is not produced by factors other than the effect under study. The presence of spurious association, due for example to the influence of extraneous variables, is called *confounding* as it tends to confound our reading and to bias our estimate of the effect studied. Conceptually, therefore, we can say that X and Y are confounded when there is a third variable Z that influences both X and Y; such a variable is then called a "confounder" of X and Y.

As simple as this concept is, it has resisted formal treatment for several decades, and for a good reason: The very notions of "effect" and "influence", relative to which "spurious association" is to be defined, has resisted mathematical formulation. The empirical definition of effect as an association that would prevail in a controlled randomized experiment, cannot easily be expressed in the standard language of probability theory, because that theory deals with static conditions, and does not permit us to predict, even from a full specification of a population density function, what relationships would prevail if conditions were to change, say from observational to controlled studies. Such predictions require extra information, in the form of causal or counterfactual assumptions [Greenland and Robins 1986; Wickramaratne and Holford 1987], which is not discernible from density functions.

These difficulties notwithstanding, epidemiologists, biostatisticians, social scientists and economists[1] have made numerous attempts to express confounding in statistical terms, partly

*Technical Report R-256, July 1998. Draft copy – comments welcome.

[1]In econometrics, the problem is mirrored by decades-long difficulties in defining "exogeneity", [Engle et al., 1983; Leamer, 1985; Aldrich 1993] which stands essentially for *no confounding.* According to Aldrich (1993) and Imbens (1997) the difficulties are still unsettled.

because statistical definitions, free of theoretical terms of "effect" or "influence," can be expressed in conventional mathematical form, and partly because such definitions may lead to practical tests of confounding, so as to alert investigators whenever adjustment is necessary. These attempts have converged on the following criterion:
Associational criterion: X and Y are not confounded if every variable Z that is not affected by the exposure variable X is either:

(U_1) unassociated with X, or

(U_2) unassociated with the outcome Y within strata of X.

This criterion, with minor variations or derivatives, can be found in almost every epidemiology textbook [Rothman, 1986; Schlesselman, 1982; Rothman and Greenland, 1998] and in almost every article dealing with confounding (e.g., [Miettinen and Cook, 1981; Greenland et al., 1997; Weinberg, 1993]). In fact, the criterion has become so deeply entrenched in the literature, that authors (e.g., Gail, 1986; Hauck et al., 1991; Steyer et al., 1996) often take it to be *the definition* of no confounding, forgetting that ultimately confounding is only so far useful as it tells about effect bias.[2]

The purpose of this paper is to bring to the attention of investigators several basic limitations of the associational criterion. We will show that the associational criterion does not ensure unbiased effect estimates, nor does it follow from the requirement of unbiasedness. After demonstrating, by examples, the absence of logical connections between the statistical and the causal notions of confounding, we will define a stronger notion of unbiasedness, called *stable unbiasedness*, relative to which a modified statistical criterion will be shown necessary and sufficient. The necessary part will then yield a practical test for stable unbiasedness which, remarkably, does not require knowledge of all potential confounders in a problem. Finally, we will argue that the prevailing practice of substituting statistical criteria for the effect-based definition of confounding is not entirely misguided, because stable unbiasedness is in fact what investigators have been and should be aiming to achieve, and stable unbiasedness is what statistical criteria can test.

2 FORMAL UNDERPINNING

To facilitate the discussion, we shall first cast the causal and statistical definitions of no confounding in mathematical forms.[3]

Definition 1 (no confounding: causal definition)
Let M be a causal model of the data-generating process, that is, a formal description of how the value of each observed variable is determined. Denote by $P(y|do(x))$ the probability of

[2] Hauck et al. (1991) dismiss the effect-based definition of confounding as "philosophic" and consider a discrepancy between two statistical criteria to be a "bias". Grayson (1987) even goes as far as stating that the change-in-parameter method, a derivative of the associational criterion, is the only fundamental definition of confounding (see [Greenland et al., 1989], for critiques of Grayson's position).

[3] For simplicity, we will limit our discussion to unadjusted confounding; extensions involving measurement of auxiliary variables are straightforward and can be found elsewhere [Pearl, 1995; Greenland et al., 1997].

the response event $Y = y$ under the hypothetical intervention $X = x$, calculated according to M. We say that X and Y are not confounded in M if and only if

$$P(y|do(x)) = P(y|x) \tag{1}$$

where $P(y|x)$ is the conditional probability generated by M.[4]

We bear in mind that the operator $do(x)$, hence effect estimates and confounding, must be defined relative to a specific causal, or data-generating model M; as these notions are not statistical in character and cannot be defined in terms of joint distributions. A given joint distribution $P(x, y)$ may be compatible with many different effect probabilities $P(y|do(x))$, depending on the model that generates $P(x, y)$. For example, we may have $P(y|do(x)) = P(y|x)$ in a model where X is an unconfounded cause of Y, $P(y|do(x)) = P(y)$ in a model where X has no causal effect on Y, while both models generate the same $P(x, y)$.

Definition 2 (no confounding: associational criterion)
Let T be the set of variables in a problem that are not affected by X. We say that X and Y are not confounded in the presence of T if every member Z of T satisfies:

(U_1) *Z is not associated with X, (i.e., $P(x|z) = P(x)$), or*

(U_2) *Z is not associated with Y within strata of X (i.e., $P(y|z, x) = P(y|x)$).*

Note that the associational criterion in Definition 2 is not purely statistical, as it invokes the predicate "affected by" which is not discernible from probabilities, and rests on judgmental knowledge about causal relationships in the domain. This exclusion of variables that are affected by treatments (or exposures) is unavoidable, and has long been recognized as a necessary judgmental input in every analysis of treatment effect, in both observational and experimental studies [Cox, 1958, p. 48]. We shall assume throughout that investigators possess the substantive knowledge required for reliably distinguishing between variables that are affected by the treatment X and those that are not. We shall see later that this distinction alone is not sufficient for establishing even the weakest test for confounding; knowledge of whether variables affect the outcome Y will also be necessary.

[4]Readers unfamiliar with the $do(x)$ operator [Pearl, 1995] may interpret $P(y|do(x))$ as the conditional probability $P^*(Y = y|X = x)$ corresponding to a controlled experiment in which X is randomized. This probability can be calculated from a causal model M either directly, by simulating the intervention $do(X = x)$, or (if M is Markovian) via the G-formula [Robins, 1986]

$$P(y|do(x)) = \sum_s P(y|x, s)P(s)$$

where S stands for all variables preceding X. An equivalent definition of $P(y|do(x))$ can also be formulated using a potential response variable Y_x [Rubin, 1974] to read

$$P(y|do(x)) = P(Y_x = y)$$

3 HOW THE STATISTICAL CRITERION FAILS

We will say that a criterion is *permissive* if it fails to detect cases of confounding, and *restrictive* if it fails to detect cases of no confounding. There are several ways that the statistical criterion of Definition 2 fails to match the causal criterion of Definition 1, some are permissive and some restrictive. These will be addressed in turns.

3.1 Permissiveness Due to Individuation

The criterion in Definition 2 is based on testing each element of T individually. A situation may well be present where two confounders, Z_1 and Z_2, jointly confound X and Y and yet, each element separately satisfies (U_1) and (U_2). This may occur because statistical independence between X and individual members of T does not guarantee the independence of X and groups of variables taken from T. An attempt to remedy Definition 2 by replacing Z with T in (U_1) and (U_2) would be much too restrictive, because the set of *all* causes of X and Y, when treated as a group, would surely fail the tests of (U_1) and (U_2).

3.2 Permissiveness Due to Small-World Assumptions

This aspect of permissiveness points to the impracticality of using the associational criterion as a test for confirming cases of no confounding. To reach such confirmation, the associational criterion requires that conditions (U_1) and (U_2) be satisfied for every potential confounder Z in a problem. In practice, since investigators can never be sure whether a given set T of potential confounders is complete, the statistical criterion will fail to recognize certain cases as confounded.

This limitation implies in fact that any statistical test whatsoever is destined to be over-permissive. Since practical tests always involve proper subsets of T, the most we can hope to achieve by statistical means is a test that would reliably detect cases of confounding when criteria such as (U_1) and (U_2) are violated. This prospect, too, will turn out unachievable, as the next two subsections demonstrate.

3.3 Restriction Due to Barren Proxies

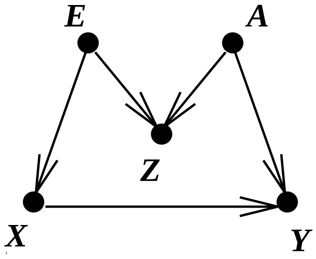
Example 1 Imagine a situation where exposure is influenced by a person's education (E), disease is influenced by both exposure (X) and age (A), while car-type (Z) is influenced by both age (A) and education (E). These relationships are shown schematically in Figure 1.

The variable car-type (Z) violates the two conditions in Definition 2 because:

1. car-type is indicative of education, hence, it is associated with the exposure variable

2. car-type is indicative of age, hence it is a associated with the disease among the exposed and the nonexposed.

Yet, in this example the effect of X on Y is not confounded; the type of car owned by a person has no effect on either exposure or disease and is merely one among many irrelevant properties that are associated with both via intermediaries. Formal analysis (e.g., [Pearl, 1995])



X	**-**	**exposure**
Y	**-**	**disease**
Z	**=**	**type of car owned by patient**
E	**=**	**education**
A	**=**	**age**

Figure 1:

establishes that, indeed, Eq. (1) is satisfied in this model, and, moreover, adjustment for Z would generally yield a biased result:

$$\sum E(Y|X = x, Z = z)P(Z = z) \neq E(Y|do(x))$$

Thus we see that the traditional criterion based on statistical association fails to identify an unconfounded effect and would tempt one to adjust for the wrong variable. This failure occurs whenever we attempt to adjust for a variable Z that is a *barren proxy*, that is, a variable that has no influence on X or Y but is a proxy for factors that do have such influence.

Readers may not consider this failure to be too serious because experienced epidemiologists would rarely regard a variable as confounder unless it is suspect of having some influence on either X or Y. Nevertheless, adjustment for proxies is a prevailing practice in epidemiology and should be done with great caution [Weinberg, 1993; Greenland et al., 1997]. To reflect this caution, the statistical criterion must be modified to exclude barren proxies from the test set T. This yields a modified criterion in which T comprises only variables that (causally) influence Y (possibly through X):

Definition 3 (no confounding: modified associational criterion)
Let T be the set of variables in a problem that are not affected by X but may affect X or Y. X and Y are said to be unconfounded by the presence of T if every member of Z of T satisfies either (U_1) or (U_2) of Definition 2.

J. Robins [Robins and Pearl, 1997] has devised an alternative modification of Definition 2 which avoids the problems created by barren proxies without requiring one to judge whether a variable has an effect on Y. Instead of restricting the set T to causes of Y, conditions (U_1) and (U_2) are replaced by the requirement that T should be composed of two disjointed subsets T_1 and T_2, such that both:

(U_1^*) T_1 is unassociated with X, and

(U_2^*) T_2 is unassociated with Y given X and T_1.

This modification of the associational criterion further rectifies the permissiveness due to individuation (subsection 3.1) because (U_1^*) and (U_2^*) treat T_1 and T_2 as compound variables. However, because T_1 and T_2 must include *all* potential confounders, we cannot conclude the presence of confounding upon testing proper subsets of T, as aspired in subsection 3.2. Thus, this criterion cannot be considered a basis for practical tests of detecting confounding.

However, the next subsection reveals a more fundamental limitation in our ability to test confounding by statistical means.

3.4 Restriction Due to Incidental Cancelations

We will present a case void of barren proxies in which the effect of X on Y is not confounded in the sense of Eq. (1) and yet it is confounded in the sense of Definition 3.

Example 2 Consider a causal model defined by the linear equations:

$$x = \alpha z + \epsilon_1 \quad (2)$$

$$y = \beta x + \gamma z + \epsilon_2 \quad (3)$$

where ϵ_1 and ϵ_2 are correlated unmeasured variables having $cov(\epsilon_1, \epsilon_2) = r$ and Z is an exogenous variable, uncorrelated with ϵ_1 or ϵ_2. The diagram associated with this model is depicted in Figure 2.

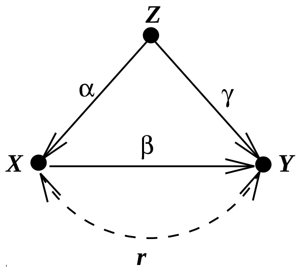


Figure 2:

It is not hard to show that the regression of Y on X gives (assuming standardized variables):

$$y = (\beta + r + \alpha\gamma)x + \epsilon$$

where $cov(x, \epsilon) = 0$. Thus, whenever the equality $r = -\alpha\gamma$ holds, the regression coefficient of X is an unbiased estimate of β, meaning that the effect of X on Y is unconfounded (no adjustment is necessary). Yet the associational conditions (U_1) and (U_2) are both violated by the variable Z, since Z is associated with X (if $\alpha \neq 0$) and with Y, given X (if γ is chosen properly, to render $\rho_{yz \cdot x}$ nonzero).

This example demonstrates that the condition of unbiasedness (Definition 1) does not imply the modified criteria of Definition 3. Hence, the associational criterion might falsely classify some unconfounded situations as confounded and, worse yet, adjusting for the false confounder (Z in our example) will introduce bias into the effect estimate.

4 STABLE UNBIASEDNESS

This discrepancy calls for a reexamination of the notion of confounding and unbiasedness as defined in Eq. (1). The reason that X and Y were classified as unconfounded in Example 2 was that, by setting $r = \alpha\gamma$, we were able to make the spurious association represented by r *cancel* the one mediated by Z. In practice, such perfect cancelation would be an incidental event, specific to a peculiar combination of experimental conditions, and would not persist when the parameters of the problem (i.e., α, γ, and r) undergo slight changes, say when the study is repeated in a different location or at different time. In contrast, the condition of no confounding found in Example 1 does not exhibit such volatility. In this example, the unbiasedness expressed in Eq. (1) would continue to hold regardless of the strength of connection between education and exposure and regardless on how education and age influence the type of car that a patient owns. We will call this type of unbiasedness *stable*, as it is robust to change in parameters, and remains intact as long as the configuration of causal connections in the model remains the same.

This distinction between stable and incidental unbiasedness behooves us to reexamine whether we should regard a criterion as over-restrictive if it misclassifies as confounded cases of incidental cancelation and, more fundamentally, whether we should insist on including such peculiar cases in the definition of unbiasedness given the precarious conditions under which Eq. (1) is satisfied. While these questions are partly a matter of choice, there is ample evidence that our intuition regarding confounding is driven by considerations of stable unbiasedness, not merely incidental one. On the pragmatic side, failing to detect situations of incidental unbiasedness should not introduce a noticeable error in observational studies because, as we have seen in Example 2, statistical tests are incapable of recognizing such cases.

Assuming that we are prepared to classify as unbiased only cases in which unbiasedness remains robust to changes in parameters, the questions remain (1) how to give this new notion of "stable unbiasedness" a formal, nonparametric formulation and (2) whether statistical criteria are available for testing stable unbiasedness. Both questions can be answered using an approach based on graphical models.

Pearl [1993, 1995] describes a graphical criterion, called the "back-door criterion", for identifying conditions of unbiasedness in a causal diagram. In the simple case of no adjustment (for measured covariates), the criterion states that X and Y are unconfounded if every path from X to Y that ends with an arrow pointing into X must also contain a pair of arrows pointing head to head. The criterion is valid whenever the missing links in the diagram represent absence of causal connections among the corresponding variables. Because the causal assumptions embedded in the diagram are so explicit, the back-door criterion has two remarkable features. First, no statistical information is needed; the topology of the diagram suffices for reliably determining whether an effect is confounded (in the sense of Definition 1) and whether an adjustment for a set of variables is sufficient for removing confounding when one exists. Second, any model that meets the back-door criterion would in fact satisfy Eq. (1) in an infinite class of models (or situations), each generated by assigning different parameters to the causal connections in the diagram.

To illustrate, consider the diagram depicted in Figure 1 of Example 1. The back-door criterion will identify the pair (X, Y) as unconfounded, because the only path ending with an arrow into X is the one traversing (X, E, C, A, Y), and this path contains two arrows pointing head-to-head at C. Moreover, since the criterion is based only on graphical relationships, it is clear that (X, Y) will continue to be classified as unconfounded regardless of the strength or type of causal relationships that are represented by the arrows in the diagram. In contrast, consider Figure 2 in Example 2, in which two paths end with arrows into X. The back-door criterion will fail to certify the effect of X on Y as unconfounded because none of these paths contain head-to-head arrows. This determination ignores the fact that under special choice of parameters $(r = -\alpha\gamma)$ the effect of X on Y may be unbiased.

To formally distinguish between *stable* and *incidental* unbiasedness, we use the following general definition:

Definition 4 *Let A be a set of assumptions (or restrictions) on the data-generating process, and let C_A be a class of causal models satisfying A. The effect of X on Y is said to be* stably unbiased *given A if $P(y|do(x)) = P(y|x)$ holds in every model M in C_A. Correspondingly, we say that the pair (X, Y) is stably unconfounded, given A.*

The assumptions often used to specify causal models are both parametric and topological. For example, the structural equation models used in the social sciences and economics are usually restricted by the assumptions of linearity, and normality. C_A in this case would consist of all models created by assigning different values to the unspecified parameters in the equations and to the non-zero entries of the covariance matrix. A more general type of assumptions emerge when we specify merely the topological structure of the diagram, but let the error distributions and the functional form of the equations remain undetermined. We now explore the statistical ramifications of this general type of assumptions.

Definition 5 *Let A_D be the set of assumptions embedded in a causal diagram D, then X and Y are* stably unconfounded *given A_D if $P(y|do(x)) = P(y|x)$ holds in every parameterization of D. By parameterization we mean an assignment of functions to the links of the diagram and prior probabilities to the exogenous variables in the diagram.*

A formal explication of the assumptions embedded in a causal diagram is given in [Pearl, 1995]. Succinctly, if D is the diagram associated with the causal model then: Every missing arrow, say between X and Y, represents the assumption that X has no effect on Y once we intervene and hold the parents of Y fixed. Every missing bi-directed link between X and Y represents the assumption that there are no common causes for X and Y, except those shown in D. Whenever the diagram D is acyclic, the back-door criterion provides a necessary and sufficient test for stable no confounding which, in the simple case of no adjustment, reduces to the nonexistence of a common ancestor, observed or latent, of X and Y.[5] Thus, we have:

Theorem 1 *Let A_D be the set of assumptions embedded in an acyclic causal diagram D. Variables X and Y are stably unconfounded given A_D if and only if X and Y have no common ancestor in D.*

[5]In the diagram of Figure 2, for example, X and Y are understood to have two common ancestors; one is Z and the second is the (implicit) latent variable responsible for the double-arrowed arc between X and Y (i.e., the correlation between ϵ_1 and ϵ_2.)

Proof. The "if" part follows from the validity of the back-door test [Pearl, 1995]. The "only if" part requires the construction of a specific model in which Eq. (1) is violated, whenever X and Y have a common ancestor in D. This is easily done using linear models and Wright's rules for path coefficients. □

Theorem (1) provides a necessary and sufficient condition for stable no confounding without invoking statistical data, as it relies entirely on the information embedded in the diagram. Of course, the diagram itself has statistical implications which can be tested [Pearl, 1988], but those tests do not specify the diagram uniquely [Pearl and Verma, 1991; Spirtes et al., 1993].

Suppose however that we do not possess all the information required for constructing a causal diagram but, instead, we know merely for each variables Z whether Z has causal influence on Y and whether Z is affected by X. The question is whether this more modest information, together with statistical data, is sufficient to qualify or disqualify a pair (X, Y) as stably unconfounded. The answer to the latter is positive.

5 OPERATIONAL TEST FOR STABLE NO CONFOUNDING

Theorem 2 *Let A_Z be the assumption that the data is generated by some (unspecified) acyclic model M, and that Z is a variable in M that is both unaffected by X and possibly affecting Y. If any of the associational criteria (U_1) and (U_2) of Definition 2 is violated, then (X, Y) are not stably unconfounded given A_Z.*

Proof. Whenever X and Y are stably unconfounded, Theorem 1 rules out the existence of a common ancestor of X or Y in the diagram associated with the underlying model. The absence of a common ancestor, in turns, implies the satisfaction of either (U_1) or (U_2), whenever Z satisfies A_Z. This is a consequence of the d-separation rule [Pearl, 1988; Greenland et al., 1997] for reading the conditional independence relationships entailed by a diagram.[6]
□

Theorem 2 implies that the traditional associational criteria (U_1) and (U_2) could be used in a simple operational test for stable no confounding, as it does not require us to know the causal structure of the variables in the domain, or even to enumerate the set of relevant variables. Finding just one variable Z that satisfies A_Z and violates (U_1) and (U_2) permits us to disqualify (X, Y) as stably unconfounded, though (X, Y) may incidentally be unconfounded in the particular experimental conditions prevailing in the study.

To the best of my knowledge, Theorem 2 communicates the first formal connection between statistical associations and confounding that is not based on the small-world assumption. It is remarkable that the connection can be formed under such weak set of added assumptions; the qualitative assumption that a variable has influence on Y and is not affected by X suffices to produce a necessary statistical test for stable no confounding.

Theorem 2 also establishes a formal connection between confounding and "collapsibility" — a criterion under which a measure of association remains invariant to the omission of certain variables.

[6] It is also an immediate corollary of Theorem 7(a) in [Robins and Pearl, 1997].

Definition 6 (Collapsibility) *Let $g[P(y,x)]$ be any functional that measures the association between Y and X in the joint distribution $P(x,y)$. We say that g is collapsible on a variable Z if*

$$E_z g[P(x,y|z) = g[P(x,y)]$$

It is not hard to show that if g stands for any linear functional of $P(y|x)$, for example the risk difference $P(y|x_1) - P(y|x_2)$, then collapsibility holds whenever Z is either unassociated with X or unassociated with Y given X. Thus, any violation of collapsibility means violation of the statistical criterion of Definition 2, hence violation of stable unbiasedness. This provides a rationale for the widespread practice of testing confoundedness by the change-in-parameter method, that is, labeling a variable Z a confounder whenever the "crude" measure of association, collapsed over the levels of Z, is not equal to the stratum-specific measures of association [Breslow and Day, 1980; Kleinbaum et al., 1982; Yanagawa, 1984; Grayson, 1987]. Theorem 2 suggests that the intuitions responsible for this practice were shaped by a quest for a stable condition of no confounding, not merely incidental one. Moreover, condition A_Z in Theorem 2 justifies a requirement made by some authors that a confounder must be a causal determinant of, not merely associated with, the outcome variable Y.

6 CONCLUSIONS

Past efforts to establish a theoretical connection between statistical associations (or collapsibility) and confounding have been unsuccessful for three reasons. First, the lack of mathematical language for expressing claims about causal relationships and effect bias has made it difficult to assess the disparity between the requirement of effect unbiasedness (Definition 1) and statistical criteria purporting to capture unbiasedness.[7] Second, the need to exclude barren proxies from consideration has somehow escaped the attention of researchers (In Example 1, $P(y|x)$ is not collapsible on Z, and still X, Y are stably unconfounded.) Finally, the distinction between stable and incidental unbiasedness has not received the attention it deserves and, as we have observed in Example 2, no connection can be formed between associational criteria (or collapsibility) and confounding without a commitment to the notion of stability. Such commitment rests critically on the conception of a causal model as an assembly of independent mechanisms which may vary independently of one another. It is only in anticipation of such independent variations that we are not content with incidental unbiasedness, but seek conditions of stable unbiasedness. The mathematical formalization of this conception has led to related notions of *graph-isomorph* [Pearl, 1988, p. 93], *stability* [Pearl and Verma, 1991] and *faithfulness* [Spirtes et al., 1993] which assist the elucidation of causal diagrams from sparse statistical associations. The same conception has evidently been shared by authors who aspired to connect statistical criteria with confounding.

[7] The majority of papers on collapsibility (e.g., Bishop, 1971; Whittemore, 1978; Geng, 1992) motivate the topic by citing Simpson's paradox and the dangers of obtaining confounded effect estimates. Of these, only a handful pursue the study of confounding or effect estimates, most prefer to analyze the more managable phenomenon of collapsibility as a stand-alone target. Some go as far as naming collapsibility "nonconfoundedness" [Steyer et al., 1996]).

ACKNOWLEDGMENT

This paper began as a commentary on Sander Greenland's "Causation, Confounding and Collapsibility" (1997, submitted). Subsequent discussions with James Robin and Sander Greenland were extremely valuable. Other helpful discussants were Guido Imbens, Michelle Pearl, and Rolf Steyer. This research was partially supported by grants from NSF, AFOSR, Northrop Grumman Corporation, and Rockwell International.

References

[Aldrich, 1993] J. Aldrich. Reiersol, Geary and the idea of instrumental variables. *The Economic and Social Review*, 24(3):247–273, April 1993.

[Bishop, 1971] Y.M.M. Bishop. Effects of collapsing multidimensional contingency tables. *Biometrics*, 27:545–562, 1971.

[Breslow and Day, 1980] N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research; Vol. 1, The Analysis of Case-Control Studies.* IARC, 1980.

[Cox, 1958] D.R. Cox. *The Planning of Experiments.* John Wiley and Sons, NY, 1958.

[Engle et al., 1983] R.F. Engle, D.F. Hendry, and J.F. Richard. Exogeneity. *Econometrica*, 51:277–304, March 1983.

[Gail, 1986] M.H. Gail. Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In S.H. Moolgavkar and R.L. Prentice, editors, *Modern Statistical Methods in Chronic Diseas Epidemiology*, pages 3–18. Wiley, New York, 1986.

[Geng, 1992] Z. Geng. Collapsibility of relative risk in contingency tables with a response variable. *Journal Royal Statistical Society*, 54(2):585–593, 1992. simpson box.

[Grayson, 1987] D.A. Grayson. Confounding confounding. *American Journal of Epidemiology*, 126:546–553, 1987.

[Greenland and Robins, 1986] S. Greenland and J. Robins. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15:413–419, 1986.

[Greenland et al., 1989] S. Greenland, H. Morgenstern, C. Poole, and J.M. Robins. Re: 'Confounding confounding'. *American Journal of Epidemiology*, 129:1086–1089, 1989.

[Greenland et al., 1997] S. Greenland, J. Pearl, and J.M. Robins. Causal diagrams for epidimiologic research. Technical Report 97-204R, University of California, Los Angeles, November 1997. To appear in *Epidemiology.*

[Hauck et al., 1991] W.W. Hauck, J.M. Heuhaus, J.D. Kalbfleisch, and S. Anderson. A consequence of omitted covariates when estimating odds ratios. *Journal Clin. Epid.*, 44(1):77–81, 1991.

[Imbens, 1997] G.W. Imbens. Book reviews. *Journal of Applied Econometrics*, 12, 1997.

[Kleinbaum et al., 1982] D.G. Kleinbaum, L.L. Kupper, and H. Morgenstern. *Epidemiologic Research.* Lifetime Learning Publications, Belmont, California, 1982.

[Leamer, 1985] E.E. Leamer. Vector autoregressions for causal inference? *Carnegie-Rochester Conference Series on Public Policy*, 22:255–304, 1985.

[Miettinen and Cook, 1981] O.S. Miettinen and E.F. Cook. Confounding essence and detection. *American Journal of Epidemiology*, 114:593–603, 1981.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligence Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

[Pearl, 1993] J. Pearl. Comment: Graphical models, causality and intervention. *Statistical Science*, 8:266–269, August 1993.

[Pearl, 1995] J. Pearl. Causal diagrams for experimental research. *Biometrika*, 82:669–710, December 1995.

[Pearl and Verma, 1991] J. Pearl and T. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452. Morgan Kaufmann, San Mateo, CA, 1991.

[Robins, 1986] J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period - applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[Robins and Pearl, 1997] J. Robins and J. Pearl. Causal effects of dynamic policies, 1997. In preparation.

[Rothman, 1986] K.J. Rothman. *Modern epidemiology.* Brown Little, 1st edition, 1986.

[Rothman and Greenland, 1998] K.J. Rothman and S. Greenland. *Modern Epidemiology.* Brown Little, 2nd edition, 1998.

[Rubin, 1974] D.B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[Schlesselman, 1982] J.J. Schlesselman. *Case-Control Studies: Design Conduct Analysis.* Oxford University Press, Oxford, 1982.

[Spirtes et al., 1993] P. Spirtes, C. Glymour, and R. Schienes. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.

[Steyer et al., 1996] R. Steyer, S. Gabler, and A.A. Rucai. Individual causal effects, average causal effects, and unconfoundedness in regression models. In F. Faulbaum and W. Bandilla, editors, *SoftStat'95, Advances in Statistical Software 5*, pages 203–210. Lucius & Lucius, Stuttgart, 1996.

[Weinberg, 1993] C.R. Weinberg. Toward a clearer definition of confounding. *American Journal of Epidemiology*, 137:1–8, 1993.

[Whittemore, 1978] A.S. Whittemore. Collapsibility of multide. *Journal of the Royal Statistical Society, B*, 40(3):328–340, 1978.

[Wickramaratne and Holland, 1987] P.J. Wickramaratne and T.R. Holland. Confounding in epidemiologic studies: The adequacy of the control group as a measure of confounding. *Biometrics*, 43:751–765, 1987.

[Yanagawa, 1984] T. Yanagawa. Designing case-contol studies. *Environmental health perspectives*, 32:219–225, 1984.