

## Myth, Confusion, and Science in Causal Analysis

J. Pearl\*

*Cognitive Systems Laboratory  
Computer Science Department  
University of California, Los Angeles, CA 90024 USA*

### SUMMARY

This letter argues that the practice of conditioning on all observed covariates, recently advocated by several analysts, should be treated with great caution. Graphical methods explain why, and provide the scientific basis for principled selection of covariates.

KEY WORDS: causality , confounding , propensity score , potential-outcome , graphical models

### Introduction

In a Letter to the Editor of *Statistics in Medicine*, Ian Shrier [1], presented a question to Don Rubin (paraphrased):

Is it possible that, asymptotically, the use of Propensity Scores (PS) methods may actually *increase*, not decrease, overall bias, compared with the crude, unadjusted estimate of a causal effect?

As most students of causality know, the answer is of course, Yes; the  $M$ -graph model presented by Shrier (see also [2], p. 186, and [3]) provides a simple such example; the crude estimate in this example is bias-free, while PS methods, or any method that adjusts for the measured covariate (a collider), introduce new bias.<sup>†</sup>

In his reply [4], Rubin did not address Shrier's question. To do so would have required modeling the treatment assignment mechanism in a communicable scientific language (e.g., graphs or structural equations), which Rubin vigorously opposes. Instead, Rubin pleaded to be "puzzled" and "confused" by the terminology, by the example, and by graphs in general,

---

Contract/grant sponsor: Office of Naval Research; contract/grant number: N000-14-09-1-0665

\*Correspondence to: Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, CA 90024 USA

<sup>†</sup>For the novice, the  $M$ -structure contains the path  $X \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$  where  $X, Y, Z$  are measured and  $U_1$  and  $U_2$  are unmeasured; conditioning on  $Z$  creates spurious dependence between  $X$  (treatment) and  $Y$  (outcome), and thus would bias our estimate of the causal effect of  $X$  on  $Y$ .

and repeated his decade-long confession that he does not find graphs to be “helpful to clear thinking about the estimation of causal effects”

Hoping to illuminate Rubin’s confusion, three letters were then sent to *Statistics in Medicine*, by Shrier [5], Sjölander [6], and myself [7], in which the nature of the  $M$ -bias was explained and exemplified, and Shrier’s question repeated and demonstratively shown to yield an affirmative answer.

To no avail. In his Author Reply, Rubin [8] rejected all three letters as “based on an unprincipled and confused theoretical perspective.” Fortunately, this time Rubin understood Shrier’s question and, anxious to demonstrate the superior insight of his graph-less model, answered it with a sweeping and false NO: “to avoid conditioning on some observed covariates,... is nonscientific ad hockery.”

Not surprisingly, Rubin’s attempt to defend this answer uncovered a much deeper confusion, this time touching on the nature of confounding, the dictates of the Bayesian paradigm, and the norms of principled scientific perspectives. Given the esteemed stature of professor Rubin in statistics circles, it is important to discuss these confusions at some length, for they may be broadly held among readers of SIM.

### 1. The more the better?

Rubin starts by arguing that the phenomenon of  $M$ -bias is not new; it merely reflects the well known fact that two independent variables may become dependent once we condition on a third.

Indeed, the phenomenon rests on the well known Berkson paradox [9] (see [2], p. 17, for historical background), in which two independent causes of a common effect (e.g.,  $U_1$  and  $U_2$  in footnote †) become dependent when we observe the effect; intuitively, information refuting one cause should make the other more likely. This phenomenon is in fact the basis for reading independencies in graphical models (DAG’s) where conditioning on a collider (or a descendant thereof) is treated differently from conditioning on a non-collider (see [10, 11]).

Given that the phenomenon is common and well known, the question arises why many researchers within the potential-outcome paradigm are unaware of its logical consequences, namely, that adding a covariate to the analysis may create spurious associations between treatment and outcome and this, in turns, may increase or decrease confounding bias. Paul Rosenbaum, for example, is one of many researchers who seems unaware of this possibility, for he writes: “there is no reason to avoid adjustment for a variable describing subjects before treatment” [12], p. 76.‡

Rubin tries to defend this tradition using an example in which gender is the covariate in question. This, of course, does not address Shrier’s question, because it a priori excludes  $M$ -bias from consideration. Gender is a variable that is exogenous in most studies of interest and, hence, cannot be modeled as a collider and cannot exhibit  $M$ -bias. With this in mind, Rubin’s logic reads as follows: since it is silly not to adjust for gender in observational studies, it must

---

‡My survey of the propensity score literature has revealed that this belief is shared by almost all authors who do not use graphs. Alerting them to its fallacy has resulted in thankful acknowledgments, with the exception of a few hard-liners whom I prefer not to name.

be “nonscientific” to refrain from adjusting for any measured covariate. I do not know many readers who would be persuaded by this logic, but one can never predict the psychological ripple of a bad example.

## 2. Is $M$ -bias rare?

Rubin’s next claim is that  $M$ -bias is a rare phenomenon, like “trying to balance a multidimensional cone on its tip.” This is factually wrong.  $M$ -bias is not a phenomenon that depends on finely-tuned, “hoped-for compensating imbalances” as caricatured by Rubin but, rather, a structural property, persisting no matter what parameters are assigned to the various associations in the model.<sup>§</sup>

While finding a pure  $M$ -structure, totally free of bias, may indeed be rare in practical studies (not unlike the rarity of finding any conditional independence,) cases containing local  $M$ -structures are abound. For example, every time we condition on a variable that is not causally related to both treatment and outcome but merely associated with the two, we may introduce an  $M$ -bias. This is because such a variable may be an indicator of several unobserved factors, some affecting treatment and some affecting outcome; by conditioning on this variable, we induce associations among those two types of factors. In doing so, we would not know if we increased, or decreased, overall bias.

As a curious and illuminating example, many of the covariates listed in Rubin’s account of the US tobacco litigation [13] fall into this category; they have no causal effect on smoking habits or on lung diseases, and yet they are statistically related to both, and are selected therefore for adjustment by astute researchers. To witness, the first covariate that appears in Rubin’s table (p. 28) is “seat-belt usage.” Obviously, seat-belt usage has no causal effect on smoking or lung diseases; it is merely an indicator of a person’s attitudes toward societal norms as well as safety and health related measures. Some of these attitudes may affect smoking habits, and some may affect susceptibility to lung diseases. If we have good reasons to believe that these two types of attitudes are marginally independent, we have a pure  $M$ -structure on our hand. But even if marginal independence does not hold precisely, conditioning on “seat-belt usage” is likely to introduce spurious associations, hence bias, and should be approached with caution. Cogently, Rubin’s sweeping advice “to condition on all observed covariates” is likely to be harmful if applied uniformly to the covariates that he himself lists in the US tobacco litigation study.<sup>¶</sup>

---

<sup>§</sup>Take two independent variables, say the outcomes,  $U_1$  and  $U_2$ , of two fair coins, and a third variable,  $Z$ , that depends on both, say a bell that rings with probability  $p = 0.75$  when  $U_1 = U_2$ . If we insist on estimating the conditional probabilities  $P(u_1, u_2|Z = 0)$  and  $P(u_1, u_2|Z = 1)$  from finite sample, it would be a “hoped-for” miracle indeed to discover that  $U_1$  and  $U_2$  are independent. However, given our prior knowledge about the science of coins, the independence of  $U_1$  and  $U_2$  ceases to be “hoped-for” and becomes structural; it can be assumed a priori without taking any measurement of  $Z$ . Even the strictest Bayesian would forgive us in this case for refraining from conditioning on  $Z$  or, for that matter, ignoring measurements of  $Z$  altogether.

<sup>¶</sup>Other listed covariates in this category are: “Insurance type,” “whether Doctor ever told having diabetes,” “whether reported suffering from arthritis,” “membership in clubs,” “home ownership,” and more.

### 3. Bayesianism versus Dogmatism

Rubin's last defense of indiscriminate adjustment appeals to Bayesian philosophy: "The Bayesian paradigm" says Rubin, "essentially directs us to condition on all observed values. To avoid conditioning on some observed covariates,... is neither Bayesian nor scientifically sound but rather it is distinctly frequentist and 'nonscientific ad hockery.'"

I differ with Rubin on this interpretation of Bayesianism, and, naturally, on his conception of principled scientific methodology. While the Bayesian paradigm teaches us indeed that one should not ignore the prior knowledge in our possession and the variables that we can observe, it does not license us to blindly condition all probabilities on those observations. Instead, it instructs us to think carefully if conditioning would advance us towards the quantity we wish estimated, or away from that quantity.

In causal analysis, Bayesianism actually directs us to refrain from conditioning on certain variables. This occurs when prior scientific knowledge informs us that conditioning would bias our estimates. Perhaps the most familiar example of such warning is the classical prohibition against adjusting for variables that lie on the causal pathway between treatment and outcome [14], p. 48. Here, qualitative knowledge about causal relationships, inexpressible in the language of statistical associations warns us to ignore available measurements. The  $M$ -structure constitutes another example in this class; again, knowledge about causal relationships, depicted graphically, warns us that adjusting for available measurements (e.g., seat-belt usage) would be inappropriate for estimating a causal effect.

There are of course problems where it is appropriate to condition on the collider of an  $M$ -structure. For example, if we merely wish to predict whether a given person is a smoker, and we have data on the smoking behavior of seat-belt users and non-users, we should condition our prior probability  $P(\textit{smoking})$  on whether that person is a "seat-belt user" or not. Likewise, if we wish to predict the causal effect of smoking for a person known to use seat-belts, and we have separate data on how smoking affects seat-belt users and non-users, we should use the former in our prediction. Such class-specific causal effects could be estimated, for example, by conducting randomized clinical trials among seat-belt users and non-users, or, in observational studies, by finding a set of covariates  $S$  that renders the class-specific causal effect among seat-belt users and non-users identifiable.<sup>||</sup> However, if our interest lies in the average causal effect over the entire population, then there is nothing in Bayesianism that compels us to do the analysis in each subpopulation separately and then average the results. The class-specific analysis may actually fail if the causal effect in each class is not identifiable. This is precisely what happens in  $M$ -structured cases; the causal effect in each sub-population is not identifiable, while the overall causal effect is. By blindly conditioning on the collider of an  $M$ -structure one does not estimate the causal effect in the corresponding subpopulation but, rather, some spurious association that may have nothing to do with cause or effect. Principled researchers, both Bayesians and non-Bayesians, should shun such methodology as "nonscientific ad hockery."

But the  $M$ -structure is but the simplest toy example where uncritical conditioning may lead to increased bias. In a complex multivariable observational study there may be several

---

<sup>||</sup>Identifiability is a more general concept than "strong ignorability," permitting a bias-free estimation of the target quantity by any means, not necessarily through adjustment. Complete criteria for the identifiability of class-specific causal effects is given in Shpitser and Pearl [15].

covariates embedded in a complex network of relationships which act like colliders in local  $M$ -structures; conditioning on these covariates may increase or decrease bias in a way that is not totally predictable without a detailed scientific model of the problem at hand.

The source of Rubin's confusion lies in refusing to accept  $M$ -structures, and graphs in general, as a legitimate representation of scientific knowledge, from which one can tell a priori whether bias may or may not be created by adjustment. The only notation that Rubin accepts is the one based on ignorability type equations. Unfortunately, this notation discourages a serious examination of prior scientific knowledge available to researchers and, consequently, it does not allow quick determination of whether conditioning is admissible. For example, to express the common knowledge that seat-belt usage ( $Z$ ) has no effect on smoking ( $S$ ) or outcome ( $Y$ ), which in the graphical language would be represented as missing arrows from  $Z$  to  $S$  or  $Y$ , would be written

$$S_z(u) = S(u) \quad \text{and} \quad Y_{zs}(u) = Y_s(u)$$

in the restricted potential-outcome language used by Rubin. To further express the assumption that factors causing seat-belt usage are mutually independent, ignorability type expressions are needed, which represent independencies among counterfactual variables.

While it is possible to translate the knowledge conveyed by a graph into formulae in potential-outcome language (see [2], pp. 98-102) the resulting mathematical expressions are so far removed from the way people communicate scientific knowledge that researchers in Rubin's camp have simply given up on representing such knowledge. This leaves them unable to tell whether "ignorability" holds or does not hold in a given problem. For many in the graph-less camp, the notion of "ignorability" is viewed as a miracle to be hoped for, not a condition that one can confirm or disaffirm from scientific knowledge, nor is it a target to be achieved by careful selection of covariates.\*\* In contrast, graphical methods permit researchers to understand what conditions covariates must fulfill before they eliminate bias, what to watch for and what to think about when covariates are selected and, not the least important, whether we have the knowledge needed for principled covariate selection.††

In general, causal inference is orthogonal to the Bayesian-frequentist debate. Berkson's paradox, the basis for  $M$ -bias, can be demonstrated using both frequency analysis and subjective human judgment. Once a Bayesian philosopher accepts and respects causal information as a legitimate component of one's prior knowledge, the Bayesian paradigm would protect one from following Rubin's advice to condition on all available measurements.

This discussion does not negate, of course, any of the teachings of the potential-outcome framework, which was shown ([2], pp. 228-234) to be a limited-vocabulary, yet mathematically equivalent version of the structural causal model that my colleagues and I have developed; a

---

\*\*A recent confession reads: "[ignorability] assumptions are usually made casually, largely because they justify the use of available statistical methods and not because they are truly believed" (workshop proceedings, unpublished). I am yet to find a single article that uses what Rubin [13] calls "the assignment mechanism,"  $P(W|X, Y(0), Y(1))$ , to determine whether "strong ignorability" holds in a given problem (see [2], pp. 341-4).

††Equally puzzling is Rubin's new definition of "ignorability," a word he helped coin. In his concluding comments, he states: "'strong ignorability' is defined to be conditional on all observed covariates." Now, what if bias can be eliminated by conditioning on a subset of the observed covariates, and not the whole set. Are we instructed then to refrain from naming the treatment "strongly ignorable given the subset?" Or should we perhaps rename it "wisely ignorable"? Frankly, I prefer the adjective "de-confoundable" over "ignorable," it is far more informative and definitely more resilient to surprising re-definitions of the kind introduced by Rubin.

theorem in one is a theorem in the other. It merely points to fallacies that are likely to emerge from an opaque notational system that has been insulated by a sub-culture of exclusivism at the expense of conceptual clarity and mathematical precision. It also points, in contrast, to insights obtained from modern, principled, and more embracing approaches to causation, ones that deploy formal and transparent representations of the science behind each problem and thus permit the mathematical analysis of confounding bias and covariate selection, liberated from myths and dogmas.

Rubin will do well to expand the horizons of his students with some of the tools that his admirers now deem illuminating.

### Acknowledgment

The author is indebted to Sander Greenland, Marshall Joffe, Thomas Richardson, and Arvid Sjölander for valuable comments on an earlier draft.

### REFERENCES

1. Shrier D. Letter to the editor. *Statistics in Medicine* 2008; **27**:2740–2741.
2. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press: New York, 2000. 2nd Edition, forthcoming July 2009.
3. Greenland S, Pearl J, Robins J. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**(1):37–48.
4. Rubin D. Author’s reply (to Ian Shrier’s Letter to the Editor). *Statistics in Medicine* 2008; **27**:2741–2742.
5. Shrier D. Letter to the editor: Propensity scores. *Statistics in Medicine* 2009; **28**:1317–1318.
6. Sjölander A. Letter to the editor: Propensity scores and  $m$ -structures. *Statistics in Medicine* 2009; **28**:1416–1420.
7. Pearl J. Letter to the editor: Remarks on the method of propensity scores. *Statistics in Medicine* 2009; **28**:1420–1423.
8. Rubin D. Author’s reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine* 2009; **28**:1420–1423.
9. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biomet. Bull.* 1946; **2**:47–53.
10. Pearl J. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann: San Mateo, CA, 1988.
11. Lauritzen S. *Graphical Models*. Clarendon Press: Oxford, 1996.
12. Rosenbaum P. *Observational Studies*. Second edn., Springer-Verlag: New York, 2002.
13. Rubin D. The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* 2007; **26**:20–36.
14. Cox D. *The Planning of Experiments*. John Wiley and Sons: NY, 1958.
15. Shpitser I, Pearl J. Identification of conditional interventional distributions. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, Dechter R, Richardson T (eds.). AUAI Press: Corvallis, OR, 2006; 437–444.