The treatment of causality in quantitative sociology: Deconfounding in *Demography* from 1995 to 2020

Jesse E. Shircliff^a and J. Tom Mueller^b

^aDepartment of Sociology and Anthropology, Utah State University, Logan UT, 84322; ^bDepartment of Geography and Environmental Sustainability, University of Oklahoma, Norman OK, 73019

Abstract

The practice of causal inference experienced significant growth throughout the social sciences in the past twenty-five years. However, it is unclear how these advances have been incorporated into quantitative sociology. To address this, we reviewed publications in *Demography* in six 5-year waves from 1995 to 2020. We assessed how authors describe the process of deconfounding statistical models. We found that the treatment of confounding factors was highly varied. Inconsistency in the descriptions of variable selection made it difficult to create a reliable quantitative measure. Still, there was minimal evidence that authors coalesced around the improvements in causal inference, at least in how they described control variables. While alignment between demography and improvements in causal inference would help consolidate and improve the science, this alignment is not evident. We explain some benefits of greater consistency and address how peer-review journals could affect the rationale behind variable selection in quantitative sociology.

Keywords: Model specification; Control variables; Systematic review; Deconfounding; Causal inference

1 Introduction

Since the 1990s, conceptual and technical improvements in causal inference literature have opened doors for the study of cause and effect from observational data (Cinelli et al. 2021; Pearl 2009; Wunsch and Gourbin 2020). For quantitative sociologists and scientists in similar fields like demography, the growth of cause-and-effect literature has broad implications for how scientists should specify statistical models using observational data (Lundberg et al. 2021). In fact, some argue that causal inference literature reached a new milestone in the past several decades by creating new, accessible techniques for causation, most of which revolve around the process of deconfounding—the process of eliminating spurious correlation between the independent and dependent variables (Pearl 2009). This milestone is evident in numerous attempts to establish a common language of model building in social sciences over the last 30 years (Greenland and Robins 1988; Moffitt 2005; Lundberg et al. 2021). In this paper, we ask: given the improvements and availability of causal literature, how have quantitative sociologists incorporated causal reasoning into their models over the same period?

We focus on how authors select variables to address deconfounding to understand whether the explanation of causal relationships in statistical inference has changed over time in quantitative sociology. In our analysis, we target quantitative sociology through the journal *Demography*. *Demography* is a leading sociological journal and the flagship journal of the Population Association of America. Although authors in *Demography* are not exclusively sociologists, demography as a field incorporates many sociological concepts (Caldwell 1996), is dominated by those working in population centers aligned with sociology departments, and the majority of demographic training occurs within or alongside sociology programs. Due to this, we argue that

the journal *Demography* serves as a valuable reflection of trends within the quantitative arm of sociology more broadly.

To look for a change in how authors address deconfounding, we systematically reviewed the description of control variables in the journal *Demography* through six volumes, sampled every fifth year from 1995 to 2020. As a leading sociological journal, it is understood that the content of *Demography* can be examined to spot major trends (Krapf et al. 2016; Teachman et al. 1993; Watkins 1993). We focus on the description of controls—variables included within a model beyond the independent variable of interest—because deconfounding is necessary for predicting causal relationships, often regardless of methods. Based on our results, authors in *Demography* have not embraced a consistent strategy or language of deconfounding. There was just a slight improvement in the description of control variables over the decades, but this process was typically under-articulated and even contradictory to causal inference literature. This trend poses the issue of latent bias in existing models and ambiguity regarding past research. We discuss how these missteps may cause quantitative sociology to lag behind other social science disciplines. Last, we recommend new approaches and protocols for authors, reviewers, and journal editors who seek a more consolidated approach to deconfounding.

2 Causal inference and deconfounding

Deconfounding is the process of eliminating factors that confound statistical measures (Cinelli et al. 2021). In statistical inference, the factors of interest are typically an independent X and a dependent Y , while a “confounder” is the factor that influences both X and Y , causing a spurious relationship to appear between the two (Gangl 2010). For example, when considering whether educational attainment (X) has a direct effect on earnings (Y), scientists must eliminate the

possibility of a third factor, such as innate ability, affecting both educational attainment and earnings (Royal Academy of Swedish Science 2021). Randomized control trials are largely considered the “gold standard” of causal inference because randomization, in theory, will evenly distribute extraneous factors like innate ability across the observed cases. Still, experimental design is typically out of reach for social scientists because they draw on observational data. Instead, social scientists from various backgrounds have developed methods and logic to control for extraneous factors and make causal interpretations of observational data possible, including the use of counterfactuals, natural experiments, and strictly specified statistical models (e.g., Neyman 1923/1990; Card 1990; Rubin 1974; Pearl 1995). That is, a central contribution of causal inference literature is the methods and logic used to deconfound the relationship between an X and a Y .

Previously, there was an animated scientific debate on how to deconfound. This confusion is exemplified by debates over the association between tobacco and shorter lifespans (Wunsch 2007). Tobacco was associated with poor health and disease, like cancer, as early as the 1920s. Yet scientists speculated that this association was a mere correlation. Fisher (1959), for example, suspected that there was a confounding gene that caused both tobacco use and cancer, meaning that smoking and lung cancer were spuriously correlated. Two studies in the 1950s, Wynder and Graham (1950) and Doll and Hill (1950) made use of case-control methods—that is, by comparing a group of people with cancer to similar a group of people without it, scientists could eliminate the possibility of a spurious relationship (Schlesselman 2006). Within the logic of deconfounding, cigarettes are now widely understood to cause poor health.

During the cigarette debate, scientists lacked a systematic approach to understanding cause and effect (Pearl and Mackenzie 2018). The main challenge was that no experimental data existed to test this association. Causal inference with non-experimental data remained in debate for many

years, but authors such as Rubin (1974:699) argued that observational data could be used given that confounding effects were “reduced or eliminated” between groups. The process of inferring cause and effect has been contested by differing definitions of controls and different models of cause and effect, but numerous authors have contributed to technical and theoretical improvements in this field (Gangl 2010; Moffit 2005, Pearl 2009; Schisterman, Cole, and Platt 2009; Wunsch 2007). The inclusion of control variables in regression models, twin studies, natural and quasi-experiments, differences-in-differences, propensity score matching, and regression discontinuity designs are all strategies that attempt to deconfound relationships between an X and a Y (Gangl 2010). These advanced methods still require authors to account for extraneous factors, so the process of deconfounding relationships remains a central concern for social scientists like sociologists and demographers. In the next section, we focus on how improvements in the causal inference literature are reflected in how models are described. Specifically, we focus on the process of selecting confounding variables to include in regression models.

2.1 Confounders and control variables

We argue that one way to assess the incorporation of improvements from causal inference literature into a given discipline is by analyzing how authors describe and select control variables for regression models. Here, we use *directed acyclic graphs* (DAGs) to illustrate the complexities in selecting controls. DAGs visualize relationships between variables. Each variable is represented by a node, and the pathway of an effect is indicated by an arrow. DAGs are used to specify relationships that move forward in time (i.e., relationships that are not cyclical) (Cinelli et al. 2021). DAGs have become increasingly common for observational studies and are considered to be one

of the major contributions of causal inference literature (Lundberg et al. 2021; Pearl and Mackenzie 2018).

As shown in Figure 1, only Panel A and B warrant the inclusion of C as a control variable in the model, since C affects both X and Y , and it will close the backdoor to another confounder, Z (Pearl and Mackenzie 2018). For example, when assessing the relationship between access to contraception (X) and fertility (Y), researchers should block variables that impact X and Y . In such a case, it would be appropriate to control for different age structures because age structure affects both the number of women who might use contraception and women's fecundity (Bongaarts 2017). It would be inappropriate to include all variables that may simply impact fertility (Y), unless each of these variables is a descendant from variables that also impact contraception (X), as seen in Panel B. This logic is known as the “backdoor criterion” (Pearl and Mackenzie 2018).

[Figure 1 here]

It is not necessary to condition on variables with mere association to X or Y , as found in Panel C (Figure 2). In this case, if unobserved factors influence C_1 or C_2 , then C_1 or C_2 brings noise into the model and biases effect estimates (Pearl and Mackenzie 2018). Variables which affect only X or Y are at best neutral in that they do not bring additional bias, but only in the latter case may they improve an estimation by limiting variation in Y (Cinelli et al. 2021). The fact is that without fully mapping out the causal path of confounding variables, researchers run the risk of opening the model to numerous forms of unexpected bias.

[Figure 2 here]

If the independent and dependent variables affect the proposed confounders, collider bias—a form of bias that occurs when X and Y are both causes of variable C —can distort the observed relationship, even inverting the effect or making it appear when it does not exist (Panel

D) (Cunningham 2021). An example of collider bias can be found in studies of COVID-19. When studying the effect of risk factors on the chances of contracting COVID-19, some researchers restricted their samples to those who were already hospitalized. In effect, these studies controlled—or deconfounded—their sample by holding hospitalization constant. However, this corresponds to the DAG in Panel D, where hospitalization (C) is both impacted by the risk factors that predict disease (X) and by contracting COVID-19 (Y). In turn, models that restricted their analysis to hospitalized cases can yield spurious correlations between risk factors (X) and the contraction of COVID-19 (Y) (Griffith et al. 2020).

We mention one more source of confusion that can occur in scenarios such as in Panel E, where a control incidentally acts as a pathway for X to influence Y (i.e., a mediation effect). In this case, a mediator introduced as a confounder will bias effect estimates toward zero if C is included (Gordon 2015). Whereas numerous covariates may increase the explanatory power of a model (i.e., R^2), without carefully considering how these variables affect X and Y , nonessential variables can simply result in over-controlling and invite the various errors described above (Pearl and Mackenzie 2018). For more examples of DAGs used to identify the types and possibilities of bias caused by inappropriate controls, see Cinelli et al. (2021).

In the case that researchers are simply interested in precise estimates of an outcome, additional covariates can sometimes increase the precision of models (Cinelli et al. 2021). However, the majority of quantitative sociology is not interested in simply maximizing R^2 and is instead interested in correctly evaluating the relationship between an X and a Y . Thus, our question is whether or not researchers use control variables to account for confounding factors and estimate precise relationships between two variables. This distinction is in-line with the recent work of Lundberg et al. (2021:533), who argue that sociologists should specifically define their target

quantity of interest, or “theoretical estimand,” in quantitative research before determining which estimation technique to employ. That is, to ensure both precise analysis and that the audience understands the study at hand, researchers should design and describe their statistical models more thoroughly (Lundberg et al. 2021). We apply this same logic to ask how authors describe control variables within the process of deconfounding.

Because modeling inappropriate variables can significantly bias effect estimates (Schisterman et al. 2009; Lu et al. 2021), we ask whether the improvements of causal inference literature are paralleled in how authors in *Demography* articulate the selection of confounding variables. We focus on changes in variable selection rather than methods because methods alone are insufficient for causal inference in most cases. For example, natural experiments and quasi-experimental approaches still require appropriate controls to make observations comparable across backgrounds (e.g., Bauer, Giesecke, Janisch 2019); fixed-effect models may eliminate confounders that are constant over time, but they still require time-varying confounders (Baltagi 2008); and differences-in-difference designs make assumptions that are often only passable after confounding variables are controlled (Lindner and McConnell 2019). As such, deconfounding remains central to advances in causal inference.

In quantitative sociology and demography, methodological and empirical rigor is commonly espoused. Given this fact, we should expect to see the use of causal logic throughout the study period with improvements occurring from 1995 to 2020. However, this is not the case. In the remainder of this article, we present our analysis of papers published in *Demography* from 1995 to 2020 and discuss implications and recommendations for the field.

3 Methodology and data

To evaluate the use of confounding variables, we adopted a systematic review method used in sociology and other fields (Borderon et al. 2019; Krapf et al. 2016; Mullenbach and Baker 2020; Page et al. 2021). First, we downloaded all 414 articles from the volumes of *Demography* published in 1995, 2000, 2005, 2010, 2015, and 2020, accessed through Scopus, JSTOR, and Springer. We screened articles to exclude any non-research publications such as erratum, corrections, editor's notes, etc., reducing the pool to 360. Next, we selected articles that assessed relationships using a regression model ($n=294$). Not all regressions are intended to assess an effect estimate, so we exclude regressions that were explicitly descriptive ($n=14$)—such as Massey and Singer (1995) who simply sought to accurately estimate the outcome of apprehension among migrants. The final number of articles reviewed was 280.

Rather than determining if every confounder was theoretically sound—thus avoiding issues of subject matter expertise—we scored how authors articulated their selection of variables. This approach focused on the discursive side of model specification. After a review of deconfounding literature, the two researchers read the entirety of the 1995 and 1996 volumes to devise a system to score how authors described their selection of control variables. The system was as follows: 1 = no controls selected; 2 = confounders controls but unexplained; 3 = an incomplete explanation of controls—such as saying controls affect only X or Y ; and 4 = controls fully explained as confounding variables that affect X and Y . To be clear, articles achieving a 3 may have been controlling for parent variables of Y or for back door confounders described in Panel B of Figure 1 (Cinelli et al. 2021). However, in articles receiving a 3, this was not clearly articulated. Since authors did not always describe each control variable in equal detail, articles were scored by their best level of explanation. Finally, topics were coded based on the dependent variable—a method

adopted from Krapf et al.'s (2016) review of gendered authorship in *Demography*. With the addition of a new category for economic demography, there were seven possible topics: 1. fertility, 2. family and household, 3. migration, 4. health and mortality, 5. economic demography, 6. data and methods, and 7. other/two or more outcomes.

To score the articles, one researcher (Blinded) scored every selected article in the remaining volumes from 2000 to 2020 (253). Then the second researcher (Blinded) coded a random selection of 10%—a portion similar to other research (Krapf et al. 2016)—of these articles using the same method. We then calculated the observed levels of agreement and calculated a measure of interrater reliability using Cohen's kappa. Kappa measures the level of agreement between different raters. Here, we used kappa to see if the score of control variable explanations were consistent (Gisev et al. 2013; McHugh 2012). In addition to the explanatory score, both researchers also recorded comments, usually with quotations, to justify the score. These records complemented the pattern (or lack thereof) observed by the quantitative measure. The researchers then discussed the trends and interpretational issues they observed while reading.

4 Results

The major takeaway was that there was little evidence that authors in *Demography* are clearly articulating the causal logic of their models. After coding the articles, the interrater reliability for the score of control variable explanations was lower than expected. Despite a set protocol, the researchers had just 58.33% agreement for explanatory scores and a kappa of 0.44. Meanwhile, the agreement on article topics was 72.22% agreement with a kappa of 0.68. A kappa of 0.44 is deemed suitable by some and unacceptable by others (McHugh 2012). Nevertheless, given that

there were four possible scores, this level of agreement is well above chance.¹ It is argued that a kappa ranging between 0.41 and 0.60 is sufficient for qualitative research while a kappa between 0.61 and 0.80 is substantial (Burla et al. 2008; Gisev et al. 2013). Thus, we believe that our findings represent legitimate trends within the field, albeit they should be interpreted with some caution. Nevertheless, our struggle to create a simple score for the description of controls was an important finding in itself, which we discuss below.

Considering both our scores and recorded comments, we found that the majority of the descriptions of control variables were incomplete across all years (Table 1). Just 16.8% of the articles described controls variables as variables that affect both *X* and *Y* (i.e., as a confounder). This percentage was just above the 14.3% that did not address control variables whatsoever. Of articles with incomplete explanations (60%), we found that variables were frequently selected simply because they impacted *Y*. It was not stated by authors that the control variables were selected because they were descendants of another confounding variable as shown in Panel C. Without strict specification, such variables could have allowed accidental mediation or opened paths for unobserved factors. The distribution of scores suggested that demographers use logic for variable selection that was out-of-step with best practices for deconfounding in observational social science.

[Table 1 here]

[Figure 3 here]

We did find limited improvement in how models were articulated over time. The percentage of articles with a given score is projected across years in Figure 3. The annual

¹ Cohen's kappa = (% observed agreement - % chance) / (1 - % chance) (Gisev et al. 2013). Assuming equal distribution of scores, chance agreement was 25%, which would yield a kappa of 0. Collapsing categories 2 and 3 (no explanation and some explanation), the explanatory score agreement was 72.22%, and its kappa was 0.59.

percentage of articles that fully explained controls as confounders steadily increased from 7.4% to just over 21% since 1995, and articles that did not mention or explain control variables decreased. That said, the number of articles with only some explanation remained high and continued to represent the vast majority of articles, typically between about 50% and 65% each year. In years when articles with only some explanation decreased, it was generally because articles without controls or with no explanation of controls had spiked, rather than articles with full explanations making up the difference. To be clear, it is possible that articles with incomplete explanations may have ultimately selected appropriate control variables and deconfounded the relationship of interest. However, the articles lacked the clarity needed for readers to evaluate the process of deconfounding on their own.

[Figure 4 here]

Similar to Krapf et al.'s (2016) work on gender and authorship, we were curious if trends varied across subject areas. *Demography*, although a journal reflecting a distinct way of conducting research, incorporates a wide variety of topical research traditions. Thus, different subject areas could approach the issue of deconfounding differently. Figure 4 shows the mean score of descriptions by topical area across years, from 1 – no controls used to 4 – controls fully explained as confounders. As can be seen, the results for this analysis did not tell a clear story with the exception of migration—which had a higher average score in all study years. Across the other topical areas, the average scores varied considerably from volume to volume. That said, the score for all topical areas ended the study period much closer to three than was found in 1995 (Figure 3).

The main limitation of this analysis was the relatively low kappa scores. Although the kappa was low, we believe our low kappa indicates an important finding on its own. Authors

specified models in a variety of ways throughout the study. Accordingly, control variable explanations also varied greatly and were difficult to evaluate—even for two researchers who developed a protocol through a close joint reading of two volumes (1995 and 1996). Approaching model specification differently is not an inherent issue, but this lack of consolidation suggests that the advances seen in causal inference literature were not reflected in quantitative sociology.

Few authors explicitly specified how they accounted for confounding factors at any level of detail, and instead, they relied on brief sections where several variables were introduced as “controls,” “adjustments,” “conditional variables,” “covariates,” and so on. The various terminology used to identify confounders were rarely unpacked, so the logic behind these terms was assumed to be taken for granted. Overall, it was often difficult to understand why variables were included at all. Moreover, control variables and the like were often explained in different sections of the paper. This was in stark contrast to information on the form of a model, for instance, which was typically found within methods or a methods sub-section. Control variables were sometimes described in the introduction, the literature review, or elsewhere—and often indirectly. Authors often simply relied on a suit of control variables used in other research, but simply citing past authors is neither a complete nor transparent means of selecting variables. In the end, reaching the same score for a single article was a challenge because the common way for authors to describe variables beyond the primary X and Y was to say as little as possible.

Our results, including the difficulties of agreement, point to one conclusion: in terms of confounding variables, the logic of causal inference, and its improvements over the past thirty years, was not reflected in how authors in *Demography* conceptualize and operationalize the selection of variables for their models. The logic of deconfounding and the quality of confounders being presented was inconsistent, difficult to evaluate, and often out-of-step with relevant causal

inference literature. For instance, despite the usefulness of DAGs, which have been promoted over a similar time period as our study (e.g., Cinelli et al. 2021; Pearl 2009), these tools were rarely used to illustrate the rationale behind variable selection. Articles published in *Demography* may have a consistent structure (introduction, background, methods, results, and discussion/conclusion), but confounding factors were not addressed systematically.

5 Discussion

To ask how quantitative sociologists used causal logic for model specification over a period of dramatic improvement in the causal inference literature, we evaluated the explanation of control variables published in *Demography* from 1995 to 2020. One of the major contributions of causal inference has been the techniques and language that allow researchers to deconfound relationships and avoid the various pitfalls of misspecification. However, the explanations provided by authors for their selection of control variables were generally weak or unclear. We found that the majority of articles did not specify controls in a manner that meets the criteria of deconfounding in contemporary causal inference literature (Cunningham 2021; Gordon 2015; Pearl 2009), nor is there strong evidence that quantitative sociology has coalesced around these improvements in the last 25 years. In turn, the field is leaving the calls for more rigorous deconfounding unanswered (Moffitt 2005; Wunsch 2007). Fortunately, the portion of articles that provided no explanations of controls variables decreased over time, and articles that fully articulated deconfounding steadily increased. The persistence of weakly justified variable selection nevertheless places the validity of specific X – Y relationships into question. It also suggests that quantitative sociologists have not leveraged their position as a statistical discipline to champion more rigorous and cutting-edge science.

The problems caused by weak deconfounding call for solutions similar to other conversations in methodology and epistemology in peer-reviewed research—that is, to focus on the role of authors, reviewers, and editorial teams in the publication process. An overreliance on p-values and the 0.05 threshold, for example, led to promoting a better understanding of the p-value itself and alternative measures of statistical confidence for authors (Halsey et al. 2015). Such discussions generated changes in journal policy and guidelines, including journals that outright banned the use of p-values, while others, such as *Demographic Research* and *The American Statistician* released statements on using and interpreting the measure (Bijak 2019; Wasserstein and Lazar 2016). Since p-values help scholars understand the variation of a statistical estimate, and the importance of including confounders is to ensure those statistical estimates honestly reflect the relationship in question, the treatment and justification of control variables is equally vital as the correct treatment of p-values.

Better, more consistent deconfounding will depend on whether authors are transparent about their research decisions, which relates to how editorial teams foster clear model specification. Controls variables should never be considered self-evident, but we found that authors in *Demography* were much more likely to describe how a variable was measured than why it was included in their model. Numerous controls could increase the precision of an estimate, but the majority of quantitative sociological scholarship is interested in estimating the effect of an X on a Y , not simply predicting an outcome with the highest level of accuracy. To estimate precise *relationships*, it is integral that confounders are addressed appropriately—meaning that control variables are selected because they affect X and Y , either directly or through a backdoor. Authors have a responsibility to explain this selection so that research audiences can understand if the theoretical orientation is focused on deconfounding. Authors should consider taking advantage of

DAGs either in the formulation of their models, or as a visual key in the paper. The important point is to avoid simply implying that control variables are appropriate or selected because they are conventional.

Editorial teams, if they wish to foster these types of changes, have an important role to play in the peer-review process. Editorial teams can motivate authors to unpack confounders through journal guidelines and protocols that relate to model specification. Likewise, reviewers should closely examine and encourage the articulation of confounding variables during the review process. Most peer-reviewed journals already expect authors to explain and address the assumptions that shape their research and analysis. We would argue the rationale for control variable inclusion is one of these key assumptions.

The strength of our paper is to show that while the causal inference literature has expanded substantially in recent decades, the process of deconfounding and the corresponding variable selection of a flagship journal for quantitative sociology has not. In fact, control variable selection remains difficult to assess, unclear, and out-of-step with these improvements. We temper our results by suggesting that quantitative sociologists may have incorporated changes that are outside the scope of this study. For example, we did not look for a general increase in causal models or techniques, such as more difference-in-difference models or the use of quasi-experimental data. Further, there are certainly quantitative sociologists who would not publish in *Demography* and are thus not represented here. By looking at different methods or journals, we may have found a different story. However, we have focused on an issue that had clear importance and room for improvement throughout the period—the selection of control variables—in a journal we argue is reflective of quantitative sociology broadly: *Demography*. Our results suggest that authors and editorial teams alike may want to step towards a more consistent treatment of causality by

improving model specification and causal reasoning practices. Until these changes are made, the ambiguity in this aspect of deconfounding ultimately hinders future scholarship and suggests the likelihood of latent bias of indeterminate direction and size.

References

- Baltagi, B. H. (2008). *Econometric analysis of panel data* (Vol. 4). Chichester: John Wiley & sons.
- Banack, H. R., & Stokes, A. (2017). The ‘obesity paradox’ may not be a paradox at all. *International journal of obesity*, 41(8), 1162 – 1163.
- Bauer, T. K., Giesecke, M., & Janisch, L. M. (2019). The Impact of Forced Migration on Mortality: Evidence from German Pension Insurance Records. *Demography*, 56(1), 25-47
- Bijak, J. (2019). P-values, theory, replicability, and rigour. *Demographic research*, 41, 949–952.
- Bongaarts, J. (2017). The effect of contraception on fertility: Is sub-Saharan Africa different? *Demographic Research*, 37, 129-146.
- Borderon, M., Sakdapolrak, P., Muttarak, R., Kebede, E., Pagogna, R., and Sporer, E. (2019). Migration influenced by environmental change in Africa. *Demographic Research*, 41, 491–544.
- Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From text to codings: intercoder reliability assessment in qualitative content analysis. *Nursing research*, 57(2), 113-117.
- Caldwell, J. C. (1996). Demography and social science. *Population Studies*, 50(3), 305–333.
- Card, D. (1990). “The impact of the Mariel boatlift on the Miami labor market.” *Industrial and Labor Relations Review*, 43: 245-257
- Cinelli, C., Forney, A., & Pearl, J. (2020). A crash course in good and bad controls. Available at SSRN, 3689437.
- Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press.
- Doll, R & Hill, A.B. (1950). Smoking and carcinoma of the lung: preliminary report. *British Medical Journal*, 2, 739-748.
- Fisher, S. R. A. (1959). *Smoking: the cancer controversy: some attempts to assess the evidence*. Edinburgh: Oliver and Boyd.
- Gangl, M. (2010). Causal inference in sociological research. *Annual review of sociology*, 36, 21-47.
- Gisev, N., Bell, J. S., and Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3), 330–338.

- Gordon, R. A. (2015). *Regression Analysis for Social Sciences*. Routledge: New York and London.
- Greenland, S. and Robins, J. M. (1988). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology International Epidemiological Association*, 15(3), 16.
- Griffith, G. J., Morris, T. T., Tudball, M. J., Herbert, A., Mancano, G., Pike, L., ... & Hemani, G. (2020). Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature communications*, 11(1), 1 – 12.
- Halsey, L.G., Curran-Everett, D. Vowler, S.L., and Drummond G.B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, 12(3), 179 – 185.
- Lindner, S., & McConnell, K.J. (2019). Difference-in-differences and matching on outcomes: a tale of two unobservables. *Health Serv Outcomes Res Method* 19, 127–144.
- Lu, H., Cole, S. R., Platt, R. W., & Schisterman, E. F. (2021). Revisiting Overadjustment Bias. *Epidemiology*, 32(5), e22 – e23.
- Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review*, 86(3), 532 – 565.
- Krapf, S., Kreyenfeld, M., and Wolf, K. (2016). Gendered authorship and demographic research: An analysis of 50 years of *Demography*. *Demography*, 53(4), 1169 – 1184.
- Massey, D. S. and Singer, A. (1995). New estimates of undocumented Mexican migration and the probability of apprehension. *Demography*, 32(2), 203 – 213.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276 – 282.
- Moffitt, R. (2005). Remarks on the analysis of causal relationships in population research. *Demography*, 42(1), 91 – 108.
- Mullenbach, L. E. and Baker, B. L. (2020). Environmental justice, gentrification, and leisure: A systematic review and opportunities for the future. *Leisure Sciences*, 42(5–6), 430 – 447.
- Mueller, J. T., Shircliff, J. E., & Steinbaum, M. (2021). Market concentration and natural resource development in rural America. *Rural Sociology*, 0(0): 1-26.
- Neyman, J. (1923/1990). “On the application of probability theory to agricultural experiments, Essays on principles, section 9.” translated in *Statistical Science*, 5, 465-480.

- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., and Moher, D. (2021). Updating guidance for reporting systematic reviews: development of the prisma 2020 statement. *Journal of Clinical Epidemiology*, 134, 103 – 112.
- Pearl, J. (1995) "Causal inference from indirect experiments." *Artificial intelligence in medicine* 7(6):561-582.
- Pearl J. (2009). *Causality* (2nd ed.), Cambridge University Press.
- Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books: New York.
- Royal Academy of Swedish Science. (2021) Answering causal questions using observational data. *The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel*. <https://www.nobelprize.org/prizes/economic-sciences/2021/advanced-information/>.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688-701.
- Schlesselman, J. J. (2006). The emerging case-control study: Lung cancer in relation to tobacco smoking. *Preventive medicine*, 43(4), 251-255.
- Schisterman, E. F., Cole, S. R., and Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(4), 488 – 495.
- Teachman, J. D., Paasch, K., and Carver, K. P. (1993). Thirty years of *Demography*. *Demography*, 30(4), 523 – 532.
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129 – 133.
- Watkins, S. C. (1993). If all we knew about women was what we read in *Demography*, what would we know? *Demography*, 30(4), 551 – 577.
- Wunsch, G. (2007). Confounding and control. *Demographic research*, 16, 97 – 120.
- Wunsch, G., & Gourbin, C. (2020). Causal assessment in demographic research. *Genus*, 76(1), 1 – 20.
- Wynder, E.L. & Graham, E.A. (1950). Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma. A study of six hundred and eighty-four proved cases. *Journal of the American Medical Association*, 143, 329-336.

5 Table and Figures

Panel A: C is a confounder

Panel B: where C closes the backdoor to confounder Z

Figure 1: Two cases where the control variable C is specified correctly (Pearl and Mackenzie 2018).

Panel C: controls only related to X or Y

Panel D: collider bias

Panel E: bias via mediation

Figure 2: Three types of misspecification that lead to biased effect estimates (Cunningham 2021; Gordon 2015).

Table 1: Percentage by score of control explanation. Only selected articles included.

Year	Articles (n)	Deconfounding category				Total
		No controls	No explanation	Some explanation	Full explanation	
1995	27	11.1	14.8	66.6	7.4	100
2000	27	11.1	25.9	51.9	11.1	100
2005	25	4.0	16.0	64.0	16.0	100
2010	53	13.2	18.9	49.1	18.9	100
2015	72	9.7	12.5	61.1	16.7	100
2020	76	5.3	7.9	65.8	21.1	100
All years	280	8.9	14.3	60.0	16.8	100

Note: some totals rounded from 99% due to decimal places.

Figure 3: This graph shows the percentage of articles by explanatory score. Each year represents one volume.

Figure 4: This graph shows the mean explanatory score by topic, where 1 = no controls, 2 = no explanation, 3 = some explanation, and 4 = full explanation. Methods papers were dropped with just 5 observations.