Testing group differences for confounder selection in nonrandomized studies: flawed practice

Nadia Sourial MSc, Isabelle Vedel MD-MPH PhD, Mélanie Le Berre MSc PT, Tibor Schuster PhD

■ Cite as: *CMAJ* 2019 October 28;191:E1189-93. doi: 10.1503/cmaj.190085

Nonrandomized studies, including observational and quasi-experimental studies, are frequently used to determine the effect of a given exposure (e.g., a new practice, intervention or policy) on relevant outcomes in situations in which random assignment to the exposed or unexposed group is not feasible or ethical.[1] A prevalent source of bias in causal inference based on nonrandomized studies is confounding. This type of bias occurs when characteristics that are causally linked to the outcome(s) of interest are imbalanced across the study groups. A known practice to identify these potential "confounders" is to test for group imbalances statistically based on the observed study data.[2–4] Characteristics found to be significantly imbalanced are then included in the statistical models in an attempt to reduce potential confounding bias.

Although, on the surface, this practice seems reasonable, observed imbalances should not guide the selection of confounders and can in fact worsen bias in estimating the exposure effect. We review why testing group differences for confounder selection in nonrandomized studies is inappropriate, assess shortcomings in established reporting guidelines and propose solutions based on recent advances in causal inference.

Why would researchers test group differences?

When imbalances are shown in variables that are conceptually consistent with confounders, this information can help corroborate existing knowledge on possible confounders. Researchers sometimes perform hypothesis tests to "confirm statistical significance" of observed imbalances and to inform the choice of variables for adjustment. For example, on reviewing articles published in *CMAJ* in 2018 (Appendix 1, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.190085/-/DC1), we found that, among the 34 nonrandomized studies that compared 2 or more groups to assess the effect of an exposure, almost one-quarter used a form of statistical testing as a means of selecting confounders for model adjustment. Although testing group imbalances can, at times, support the decision to include variables for adjustment, it can also create confusion in situations where results do not agree with preconceptions or knowledge. Testing becomes particularly problematic when used as the primary method to inform the choice of variables for confounding adjustment.

KEY POINTS

- Using observed imbalances between study groups (e.g., exposed and unexposed) to determine variables for confounding adjustment in nonrandomized studies may misguide the selection of variables to control for in the analysis and, thus, may bias study results.
- Reporting guidelines for research are vague and, in some cases, erroneous in their direction regarding this inappropriate practice.
- Advances in causal inference offer new insights and solutions on handling confounding and should be incorporated into current reporting guidelines.

What are the pitfalls of testing group differences for confounder selection?

Limiting the search for confounders to testing for group imbalances fails to consider possible unobserved variables that are relevant in the confounding mechanism.[5] Variables that are strong predictors of the outcome but are only weakly associated with the exposure would be less likely to be selected for adjustment, resulting in uncontrolled confounding and biased results. Sun and colleagues[2] showed that bivariate screening methods were insufficient to control confounding and could exclude important variables from the multivariable analysis. Groenwold and colleagues[4] showed that lack of adjustment for a baseline characteristic that is only marginally different between groups at baseline but strongly associated with the outcome can result in significant overestimation of the effect of the exposure.

At the other extreme, testing observed group differences may inadvertently identify as confounders variables that are on the causal pathway between the exposure and outcome (so-called "mediators") or variables that are a common effect of other variables, of which at least 1 is linked to the exposure and 1 to the outcome (so-called “colliders”). There are many valid scenarios in which identifying and measuring the effect of a mediator is of interest; for example, it may be more feasible or cost-effective to intervene at the level of the mediator rather than the exposure in order to affect the outcome.[6] However, when the purpose is to measure the total effect of an exposure, adjusting for mediators or colliders can, in fact, introduce rather than remove bias in the estimated effect.[3,4,7] A more detailed review of mediation analysis is available elsewhere.[8]

Overall, relying on statistical testing for confounder selection can contribute to creating a paradox in which true confounders may not be identified and adjustment for nonconfounders may create spurious associations.[3,4] Fortunately, more appropriate methods to assess confounding exist.

What do current reporting guidelines advise?

We reviewed relevant reporting guidelines for research and assessed any guidance provided for handling group imbalances in nonrandomized studies or on confounder selection. Five guidelines were reviewed. We found that the level and appropriateness of the guidance varied (Table 1). Both the International Committee of Medical Journal Editors (ICMJE)[9] and the Statistical Analyses and Methods in the Published Literature (SAMPL) guidelines[14] offer no specific guidance and refer readers to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement.[10,11] The STROBE statement provides appropriate guidance overall, recommending deciding on potential confounders at the study planning stage and discouraging the use of

Table 1: Relevant reporting guidelines related to testing group imbalances for confounder selection in nonrandomized studies

Organization/external reporting guideline	Current guidance
ICMJE[9]	Does not include guidance on reporting of research methods including testing group imbalances for confounder selection in nonrandomized studies. Refers authors to STROBE.
STROBE[10,11]	**Item 12 (a). Describe all statistical methods, including those used to control for confounding** “Investigators should think beforehand about potential confounding factors. This will inform the study design and allow proper data collection by identifying the confounders for which detailed information should be sought.” “If groups being compared are not similar with regard to some characteristics, adjustment should be made for possible confounding variables by stratification or by multivariable regression.”
	Item 14: Descriptive data “Inferential measures such as standard errors and confidence intervals should not be used to describe the variability of characteristics, and significance tests should be avoided in descriptive tables. Also, P values are not an appropriate criterion for selecting which confounders to adjust for in analysis; even small differences in a confounder that has a strong effect on the outcome can be important.”
TREND[12]	**Item 15: Baseline equivalence — data on study group equivalence at baseline and statistical methods used to control for baseline differences** “Example (baseline equivalence): the intervention and comparison groups did not statistically differ with respect to demographic data (gender, age, race/ethnicity; $p > .05$ for each), but the intervention group reported a significantly greater baseline frequency of injection drug use ($p = .03$); all regression analyses included baseline frequency of injection drug use as a covariate in the model.”
GRADE[13]	**5.2 Factors that can reduce the quality of the evidence** *5.2.1 Study limitations (risk of bias)* “Study limitations in observational studies”: • Failure to develop and apply appropriate eligibility criteria (inclusion of control population) ○ Under- or overmatching in case–control studies ○ Selection of exposed and unexposed in cohort studies from different population • Failure to adequately control confounding ○ Failure of accurate measurement of all known prognostic factors ○ Failure to match for prognostic factors and/or adjustment in statistical analysis **5.3. Factors that can increase the quality of the evidence** *5.3.3. Effect of plausible residual confounding* “Rigorous observational studies will accurately measure prognostic factors associated with the outcome of interest and will conduct an adjusted analysis that accounts for differences in the distribution of these factors between intervention and control groups.”
SAMPL[14]	Does not include specific guidance on reporting of research methods including testing group imbalances for confounder selection in nonrandomized studies. Refers authors to STROBE and TREND.

Note: GRADE = Grading of Recommendations Assessment, Development and Evaluation, ICMJE = International Committee of Medical Journal Editors, SAMPL = Statistical Analyses and Methods in the Published Literature, STROBE = STrengthening the Reporting of OBservational studies in Epidemiology, TREND = Transparent Reporting of Evaluations with Nonrandomized Designs.

significance tests using study data in confounder selection, but we found 1 statement that seemed somewhat misleading: "If groups being compared are not similar with regard to some characteristics, adjustment should be made." The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach[13] underlines the importance of adequately selecting the variables for model adjustment as well as the risk of bias from failure to control for confounding but offers no specific guidance on the methods that should or should not be used. Surprisingly, the Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) statement explicitly promotes testing group differences as a tool to select variables for adjustment.[12]

How should confounding variables be determined appropriately?

Confounding is a fundamental concept in causal inference, an area of research that has seen major developments in recent years.[15,16] The formal definition of confounders under the causal inference framework has also been the subject of recent debates.[17] Confounding is not something that can be determined or statistically tested using data alone.[3,18] Instead, selecting the set of confounders to adjust for should be considered at the design stage using a conceptual framework based on subject matter knowledge and published evidence.[3,10]

Causal diagrams known as directed acyclic graphs (DAGs) are one among other conceptual design tools that can aid researchers in confounder selection; they have been growing in popularity.[5,7,19] These diagrams provide a visual conceptualization using arrows to represent the causal pathways involved in the exposure–outcome relation (Figure 1). The graph is "directed" because arrows are unidirectional and "acyclic" because there is no path connecting a variable back to itself. A key strength of DAGs is that, using graph theory, they can differentiate between confounders and other types of variables, like mediators and colliders, to determine the set of confounders that must be taken into account when estimating the effect of interest. This is accomplished by examining the location of variables in the causal pathway and the causal links leading into or out of, or both, these variables.

Consider a hypothetical example on the effect of a transitional care program (from hospital discharge back to the community) compared to usual care on subsequent emergency department visits (Figure 1). Suppose patients were not randomly assigned to either the transitional care group or the usual care group. We may decide to consider as confounders variables found to be imbalanced between the groups. These may be the number of patient comorbidities, patients' level of continuity of care and patients' satisfaction, for example. However, mapping out the causal pathways through a DAG, we would be able to uncover which of these variables should, in fact, be adjusted for and which should not. Using a DAG at the design stage can help decide a priori on the variables and data that need to be collected.

If, for example, patient comorbidity satisfies the conditions of a confounder, as it is a common cause of both group membership and going to the emergency department, we should conclude that patient comorbidity *should* be adjusted for. If, however, transitional care leads to better continuity of care, which, in turn, leads to decreased emergency department visits, continuity of

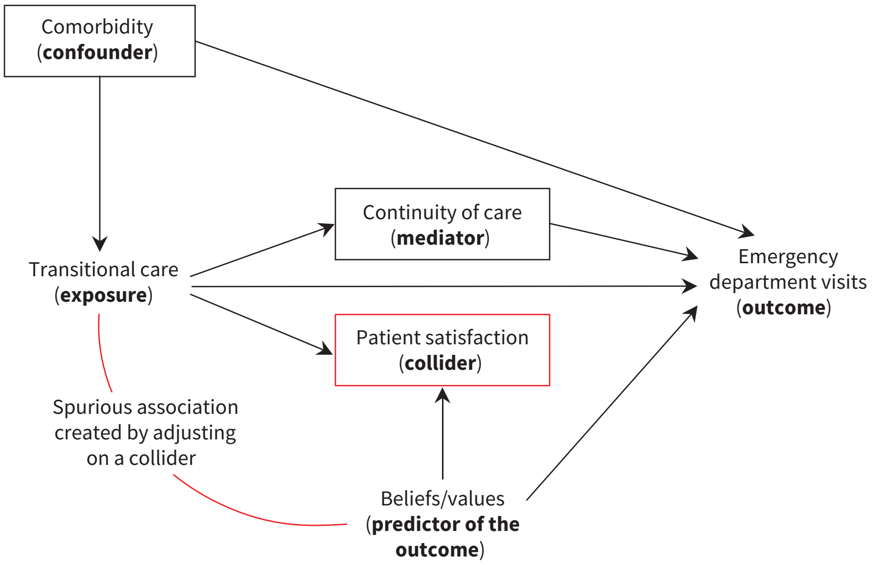

Figure 1: Hypothetical directed acyclic graph and impact of adjusting for different types of variables in the causal pathway. Boxes around a variable indicate adjustment. In this example, adjusting for comorbidity (confounder) would correctly block the spurious association between transitional care and emergency department visits due to the common cause of comorbidity. Adjusting for continuity of care (mediator) would block part of the total effect between transitional care and emergency department visits. Finally, adjusting for patient satisfaction (collider) would create a spurious association between transitional care and emergency department visits through patient beliefs and values, as indicated by the red arc connecting the colliding variables "exposure" and "predictor of the outcome."

care[24] would be a mediator, and adjusting for it would block part of the total effect of transitional care on emergency department visits. In this case we would conclude that continuity of care *should not* be adjusted for if we are interested in estimating the total effect of transitional care on emergency department visits. Furthermore, if we consider that transitional care affects patient satisfaction but that patient beliefs and values also affect patient satisfaction and emergency department visits, patient satisfaction is a common effect of both transitional care and beliefs (i.e., a collider). Adjusting for patient satisfaction would create a spurious association between transitional care and beliefs, and, as beliefs, in this example, also predict patients' decision to go to the emergency department, a spurious association would be created between transitional care and emergency department visits (through beliefs), biasing the effect of transitional care on emergency department visits. Therefore, we would conclude that patient satisfaction *should not* be adjusted for.

A tabular summary of these assessments is provided in Appendix 2 (available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.190085/-/DC1). In this hypothetical example, only adjustment on patient comorbidity would be needed to ensure that the estimated effect of transitional care on emergency department visits is free of confounding bias.

Real-world DAGs are often more complex, with a large number of interconnected variables involved in the exposure–outcome relation. Published tutorials and online tools are available to assist in developing and interpreting DAGs to tease out the set of confounders that should be adjusted for in a specific study.[20–22] It should be noted, however, that the utility of DAGs depends on the quality of the evidence on which they are based. They also require a degree of subjective judgment and therefore cannot guarantee that all true confounders will be correctly identified. Nevertheless, DAGs remain a useful tool to better understand and visualize the complex pathways involved in the exposure–outcome relation and can help more rigorously determine sources of confounding.

In addition to DAGs, another approach to confounder selection is to adjust for variables that are known (or believed) to be predictive of exposure status or of the outcome, or both. This method has been shown to be sufficient to provide adequate confounding control.[23]

For characteristics with observed group imbalances that were not considered confounders at the design stage, bias factors and confounding functions can provide useful sensitivity analyses.[24,25] These methods calculate the magnitude by which the estimated effect of the exposure is affected by a potential confounder that was not controlled for or measured. Because they incorporate the imbalance of a characteristic across exposure groups, as well as its relation with the outcome, these methods better reflect the triangular nature of confounding than imbalances with respect to exposure status alone. In their simplest form, bias factors are calculated by multiplying the difference in the prevalence of the confounder between the intervention and control groups by the effect of the confounder on the outcome. For example, Vanderweele and Arah[24] showed that if the prevalence of an unmeasured confounder is 30% higher in the intervention group than in the control group and is associated with a 52% higher risk of having the outcome, the magnitude of bias would be 0.30 × 0.52 = 0.16. In other words, by *not* adjusting for this confounder, the exposure effect would be overestimated by 16%. Confounding functions expand on bias factors by examining a range of different confounding scenarios. Their effect on the estimated effect of exposure can then be represented graphically to visualize the relation between the degree of bias and the shift in the estimated effect.[25] Details of the methods are provided elsewhere.[24,25]

Should reporting guidelines on nonrandomized studies be updated?

Given the development of new tools and methods for confounder selection within the field of modern causal inference, an update to current guidance on confounder selection in nonrandomized studies, with more uniformity across guidelines, seems warranted. We suggest that these guidelines should 1) emphasize the selection of confounders at the design stage through the use of DAGs or other conceptual tools to avoid inadvertently adjusting for mediators and colliders, 2) delineate the pitfalls of relying on observed data and the results of statistical tests such as *p* values, confidence intervals and univariate tests for confounder selection and 3) propose the use of sensitivity analyses, such as bias factors or confounding functions, at the analysis stage to assess the impact of unmeasured confounders. Engagement with end-users and authors of the reporting guidelines is also needed to formally test and revise the guidelines to ensure clarity. Endorsement of reporting guidelines by journal editors and reviewers continues to play an important role in further dispelling the practice of using observed data to inform confounding.

Conclusion

Nonrandomized studies make an important contribution to the research literature and may supply valuable evidence for practice and policy decision-making. If the evidence produced from nonrandomized studies for the purpose of causal inference is to be considered reliable, careful attention needs be paid to the quality of the methods aimed at addressing the various potential sources of bias, such as confounding arising from lack of randomization.

However, the practice of confounder selection based on statistical testing of group differences has continued to "fly under the radar." When misused, these statistical tests can mislead rather than inform on the effectiveness or safety of exposures or interventions. With advances in causal inference, we are now in a position to promote better research practice by explicitly discouraging this approach to confounder selection and endorsing more appropriate methods.

References

1. Schünemann HJ, Tugwell P, Reeves BC, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Res Synth Methods* 2013;4:49-62.
2. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996;49:907-16.

3. Brookhart MA, Stürmer T, Glynn RJ, et al. Confounding control in healthcare database research: challenges and potential approaches. *Med Care* 2010;48 (Suppl 6):S114-20.
4. Groenwold RH, Klungel OH, Grobbee DE, et al. Selection of confounding variables should not be based on observed associations with exposure. *Eur J Epidemiol* 2011;26:589-93.
5. Hernàn MA Robins JM. *Causal inference*. Boca Raton (FL): Chapman & Hall/CRC; 2018:67-104.
6. Sourial N, Longo C, Vedel I, et al. Daring to draw causal claims from non-randomized studies of primary care interventions. *Fam Pract* 2018;35:639-43.
7. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:37-48.
8. VanderWeele T. *Explanation in causal inference: methods for mediation and interaction*. London (UK): Oxford University Press; 2015:20-248.
9. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. International Committee of Medical Journal Editors; updated December 2018. Available: www.icmje.org/icmje-recommendations.pdf (accessed 2018 Aug. 26).
10. von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007;147:573-7.
11. Vandenbroucke JP, Von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med* 2007;4:e297.
12. Des Jarlais DC, Lyles C, Crepaz N, et al. Improving the reporting quality of non-randomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health* 2004;94:361-6.
13. Schünemann H, Brożek J, Guyatt G, et al., editors. GRADE handbook. GRADE Working Group; 2013. Available: http://gdt.guidelinedevelopment.org/app/handbook/handbook.html (accessed 2019 June 6).
14. Lang TA, Altman DG. Basic statistical reporting for articles published in clinical medical journals: the Statistical Analyses and Methods in the Published Literature, or SAMPL guidelines. In: Smart P, Maisonneuve H, Polderman A, editors. *Science editors' handbook*. 2nd ed. London (UK): European Association of Science Editors; 2013;175-82.
15. Pearce N, Lawlor DA. Causal inference — so much more than statistics. *Int J Epidemiol* 2016;45:1895-903.
16. Richardson TS, Rotnitzky A. Causal etiology of the research of James M. Robins. *Stat Sci* 2014;29:459-84.
17. VanderWeele TJ, Shpitser I. On the definition of a confounder. *Ann Stat* 2013;41:196-220.
18. Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001;12:313-20.
19. Pearl J. *Causality: models, reasoning and inference*. 2nd ed. New York: Cambridge University Press; 2009:65-106.
20. Textor J, van der Zander B, Gilthorpe MS, et al. Robust causal inference using directed acyclic graphs: the R package 'dagitty.' *Int J Epidemiol* 2016;45:1887-94.
21. Williamson EJ, Aitken Z, Lawrie J, et al. Introduction to causal diagrams for confounder selection. *Respirology* 2014;19:303-11.
22. DAGitty. Welcome to DAGitty! Available: http://www.dagitty.net/ (accessed 2019 June 6).
23. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics* 2011;67:1406-13.
24. Vanderweele TJ, Arah OA. Unmeasured confounding for general outcomes, treatments, and confounders: bias formulas for sensitivity analysis. *Epidemiology* 2011;22:42-52.
25. Kasza J, Wolfe R, Schuster T. Assessing the impact of unmeasured confounding for binary outcomes using confounding functions. *Int J Epidemiol* 2017;46:1303-11.

Competing interests: Nadia Sourial is financially supported by a Canadian Institutes of Health Research — Vanier Canada Graduate Scholarship.

This article has been peer reviewed.

Affiliation: Department of Family Medicine, McGill University, Montréal, Que.

Contributors: Nadia Sourial, Isabelle Vedel and Tibor Schuster contributed to the study design. Nadia Sourial, Mélanie Le Berre and Tibor Schuster drafted the manuscript. All of the authors contributed to the study conception, critically revised the manuscript for important intellectual content, approved the final version to be published and agreed to act as guarantors of the work.

Correspondence to: Tibor Schuster, tibor.schuster@mcgill.ca