European Journal of Epidemiology (2019) 34:221–222
https://doi.org/10.1007/s10654-019-00495-5

COMMENTARY

Theory meets practice: a commentary on VanderWeele's 'principles of confounder selection'

Sebastian Schneeweiss[1]

Received: 5 February 2019 / Accepted: 8 February 2019 / Published online: 6 March 2019

When I teach graduate students in pharmacoepidemiology they are well-trained by having taken multiple courses in epidemiologic methods and causal inference. I observe that many feel paralyzed when confronted with real data realizing that such data do not come with tags saying whether variables are common causes of the exposure and outcome or whether they are instrumental variables or colliders. How will they connect the concepts, rules, and exemptions they have learned studying causal inference to the reality of data? Tyler VanderWeele is to be applauded for having compassion with us who spend less time contemplating DAGs and still want to do non-experimental studies that lend themselves to causal conclusions. His pragmatic recommendations are actionable for a broad range of applications yet founded in principled considerations. I tried to put them to a test.

I am a pharmacoepidemiologist and study the safety and effectiveness of medications as they are used in routine care without the help of randomized treatment assignment. Like in other areas of epidemiology, causal conclusions are paramount as we apply our insights to patients and influence treatment decisions.

It is implicit in Dr. VanderWeele's commentary that we have a clear grip on temporality and can distinguish between pre-exposure variables and those patient characteristics measured after exposure started. As trivial as this may sound one is surprised how often this is confused, particularly when working with secondary analyses of already collected data [1]. Such mix-up occurs more frequently in case–control sampling designs nested in cohorts of patients identified within longitudinal patient databases. Staying with a cohort analysis and focusing on new users of medications with a defined inception point and covariate assessment before cohort entry substantially reduces the chance for such self-inflicted errors [2].

How do we identify the causes of exposure? Textbooks recommend "prior knowledge" or "a review of the literature" among others. Studying treatment guidelines and conversations with colleagues (better outside of major academic medical centers with their idiosyncrasies) help us to identify potential predictors of treatment choice. While this is a good starting point, predictors for treatment choice often vary substantially depending on health care systems, provider organizations and individual providers. Are we as investigators always doing a most rigorous job in identifying all potential predictors of treatment choice? Investigators studying hypoglycemic events among users of the oral anti-diabetic drugs glyburide versus glipizide inadvertently oversaw that pregnancy and gestational diabetes are strong predictors for choosing one over the other in some healthcare systems. These covariates were not pre-specified by this highly experienced investigator team [3].

Causes of the disease outcome should be identifiable by reading medical textbooks and most causes of disease should be universal. However, in a given population the strength or the observability of such risk factors of a health outcome will vary. Based on our prior knowledge, how sure are we that we have identified all risk factors of the disease and have captured them sufficiently?

In addition to these generic issues comes the complication that we usually work with secondary electronic data that were collected by operating a healthcare system, including administrative claims data and electronic health records [4]. This means we are not in control of what, when, and how to measure covariate information and we often rely on proxy variables of the constructs that we are really after. In order to identify diabetes, we may rely on a diagnostic code for diabetes. Was that code checked because a glucose test was performed to rule out diabetes? Should we therefore rather

✉ Sebastian Schneeweiss
schneeweiss@post.harvard.edu

[1] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

require at least 2 codes of diabetes within a couple of weeks, maybe plus treatment with an antidiabetic medication to be really sure? [5].

For all the reason above, the difficulty of knowing all causes of exposure and of the outcome plus the added difficulty to measure them fully, VanderWeele mentioned data-adaptive methods like high-dimensional propensity scores (HDPS) or collaborative targeted maximum likelihood estimation (CTMLE) [6, 7]. The principle idea of both is to empirically identify predictors of exposure and of the outcome in the data at hand and then prioritize variables for adjustment. HDPS is frequently used in pharmacoepidemiology as it is specifically adapted to the typical data sources we work with. CTMLE is a generalization of HDPS. It performs equally well and is seen as more opaque by decision makers [7, 8]. A simple covariate prioritization operation in HDPS [9] performs as well as a range of machine learning techniques to augment variable selection, at least in the data we typically work with [10]. It is specific to working with secondary healthcare databases that we include markers for health services utilization intensity. Diagnoses and procedures are only recorded when patients encounter the professional healthcare system and therefore utilization intensity is related to our ability to observe confounders and the surveillance for study outcomes.

Both of these data-adaptive approaches embrace the principle of VanderWeele's "common cause criterion" but in practice are often closer to his "disjunctive cause criterion" because we may not fully understand the underlying causal diagram. Like him, we are concerned about adjusting for instrumental variables (IVs) and colliders that may increase rather than decrease bias. There is strong evidence that if one cannot decide empirically whether pre-treatment variables are confounders, IVs, or colliders one is better off adjusting for them in most realistic settings [11, 12]. One can empirically screen for variables that behave like IVs and can withhold them from the adjustment strategy or, if truly lucky, they might be real IVs and one would choose an IV analysis.

As pharmacoepidemiologists we relate with regulatory agencies and payer organizations that make decisions about the availability and coverage of prescription drugs. Transparency and pre-specification of our covariate adjustment strategy are seen as critical for such high-stakes decision-making. As such we declare our selection criterion, possibly in VanderWeele's terminology, and describe our precise procedure on how we identified the covariates used for the treatment effect analysis.

In summary, Dr. VanderWheele's recommendations are much aligned with our preferred approaches in pharmacoepidemiology and outcomes research using secondary data sources.

Funding This work was funded partially by the Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, the FDA (HHSF223201710186C) and PCORI. This article reflects the views of the author and should not be construed to represent FDA's views or policies.

Compliance with ethical standards

Conflict of interests Dr. Schneeweiss is a consultant to WHISCON, LLC and to Aetion, Inc. of which he also owns equity. He is the principal investigator of Grants to the Brigham and Women's Hospital from Bayer, Vertex, and Boehringer Ingelheim unrelated to the topic of this paper.

References

1. Patorno E, Garry EM, Patrick AR, et al. Addressing limitations in observational studies of the association between glucose-lowering medications and all-cause mortality: a review. Drug Saf. 2015;38:295–310.
2. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010;19:858–68.
3. Zhou M, Wang SV, Leonard CE, et al. Sentinel modular program for propensity score-matched cohort analyses: application to glyburide, glipizide, and serious hypoglycemia. Epidemiology. 2017;28:838–46.
4. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005;58:323–37.
5. Glynn RJ, Monane M, Gurwitz JH, Choodnovskiy I, Avorn J. Agreement between drug treatment data and a discharge diagnosis of diabetes mellitus in the elderly. Am J Epidemiol. 1999;149:541–9.
6. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clin Epidemiol. 2018;10:771–88.
7. Ju C, Wyss R, Franklin JM, Schneeweiss S, Haggstrom J, van der Laan MJ. Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data. Stat Methods Med Res. 2017. https://doi.org/10.1177/0962280217744588.
8. Wyss R, Schneeweiss S, van der Laan M, Lendle SD, Ju C, Franklin JM. Using super learner prediction modeling to improve high-dimensional propensity score estimation. Epidemiology. 2018;29:96–106.
9. Bross ID. Spurious effects from an extraneous variable. J Chronic Dis. 1966;19:637–47.
10. Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. Epidemiology. 2017;28:237–48.
11. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol. 2011;174:1213–22.
12. Liu W, Brookhart MA, Schneeweiss S, Mi X, Setoguchi S. Implications of M bias in epidemiologic studies: a simulation study. Am J Epidemiol. 2012;176:938–48.