Drug Saf (2015) 38:105–107
DOI 10.1007/s40264-014-0254-8

LETTER TO THE EDITOR

Authors' Reply to Hennessy and Leonard's Comment on "Desideratum for Evidence-Based Epidemiology"

J. Marc Overhage · Patrick B. Ryan · Martijn J. Schuemie · Paul E. Stang

Published online: 16 December 2014

We appreciate Hennessy and Leonard's [1] comments on our paper and their strong support for the need to carefully characterize the performance of epidemiologic methods and analysis choices (which we collectively refer to as analyses). The work performed as part of the Observational Medical Outcomes Partnership (OMOP) is but a first step in this journey. We particularly appreciate the authors making the point that "Problematically implemented studies do not invalidate the underlying research designs, just those implementations". We fully agree with this assertion and the importance of beginning to systematically answer questions about what makes analyses problematic. We also share their belief that the empirical assessment of the performance of analyses applied to observational datasets is an essential prerequisite for understanding the reliability of any evidence developed from observational studies.

Every measurement approach is limited in precision and accuracy, and the OMOP investigators have consistently acknowledged that the performance evaluation framework we used has limitations [2]. However, without any measure of performance, we are limited to making assumptions based on our theories. Science advances primarily by testing such theories and improving them based on empiric evidence.

Ideally, we would measure the performance of analyses with a reference set of positive controls with known effect sizes and negative controls without an effect. Unfortunately, such a collection does not exist so we developed a reference set based on a systematic evaluation of multiple knowledge sources (including literature, product labeling, and a systematic review) that provide insight into others' assessment of a causal relationship between selected exposures and outcomes [3]. The evidence supporting these controls falls short of the desired, but inherently unknowable, gold standard, and is further limited by the fact that most positive controls may have been known to clinicians before our data were generated [4]. It is important to note that we relied on positive controls only for measures of discrimination, while estimates of error and calibration are based only on negative controls which suffer less from these shortcomings.

Simulated data, whether fully synthetic or created by injecting signal into real data, is another approach to creating a 'gold standard' [5]. It suffers from its own limitations, including the validity of the assumptions and the degree to which it reflects the relevant complexities of the real world. Experiments based on either simulated or real-world data alone are unlikely to be sufficient given their limitations, and offer complementary data on empirical performance. We measured the performance of analyses using fully synthetic data and obtained essentially the same results as when using the real data [6].

Despite these limitations, it is clear to us that empirically measuring performance of analyses for a particular question on a specific dataset (not generic performance as a one-size-fits-all solution) and using the empirical operating characteristics to calibrate the result is essential. Expert opinion and subjective arguments about theoretical beliefs around potential bias are not a sound foundation on which to develop evidence that is intended to be used to inform medical decisions, whether at the population or individual patient level.

As is the case for controls, there is no gold standard for the definition of health outcomes (HOIs).

J. M. Overhage (✉)
Siemens Medical Solutions, Malvern, PA, USA
e-mail: marc.overhage@siemens.com

P. B. Ryan · M. J. Schuemie · P. E. Stang
Jansen Research and Development, Titusville, NJ, USA

While some argue that medical record review is an authoritative approach, it is limited by the fact that the record is only an imperfect reflection of the patient clinical state, is flawed in organization, and human reviewers routinely commit errors in identifying data in the record [7–11]. We may be able to incrementally improve the classification of outcomes but it is unclear whether it is sufficient to meaningfully improve effect estimates. The current practice of undertaking source record verification on a small sample of identified cases to estimate a positive predictive value (PPV) may have limited value because you really need to know sensitive and specificity to properly assess misclassification, but these are rarely measured [12–14]. Furthermore, probably the most important thing to know is how much misclassification is conditional on the exposure on which source record verification can shed little light. Rather than assuming that an outcome definition has no misclassification error, we would encourage approaches to quantify the error and incorporate its potential effect into the overall effect estimates.

The HOI definitions the OMOP used were derived based on two independent, comprehensive literature reviews [15]. For each outcome, we evaluated the performance of analyses using multiple alternative definitions, and the results suggest that further work is needed to properly optimize the sensitivity–specificity tradeoff that is required in the outcome definition process [16]. Similarly to the controls, we welcome and encourage additional development of systematic, evidence-based approaches to defining and characterizing the empirical performance of alternative HOI definitions.

Concerns about problematically implemented studies and how we know when they are problematic was the focus of OMOP's work. OMOP's results highlight the very real possibility that even well-intentioned, well-trained investigators cannot, with confidence, choose a method and analysis choices that will reliably yield the 'correct' result in a specific study. In addition to the variability evident in the literature, we documented the variation in choice of analyses for the same epidemiologic question [17]. While we agree that "there are sound theoretical bases for the epidemiologic designs commonly used" we believe that all theory can and should be experimentally tested and that, by examining the vast majority of commonly employed analyses, we have identified a variety of areas in which the results of these experiments are not consistent with the theory. While not every approach for mitigating the effects of confounding were tested, a number were evaluated in the OMOP experiments, including a variety of self-controlled designs, case-control designs, and new-user cohort designs using propensity score adjustment.

When theory and experiment disagree, we need to examine the experimental design and execution to identify potential flaws, but we must also reassess the theory. The experiments that the OMOP investigators and others have performed have limitations, but the direction and consistency of the findings across analyses, databases and outcomes definitions suggest that some common assumptions on which the theory is based may not be valid. The work performed as part of the OMOP is but a first, imperfect, but informative step in the odyssey that the epidemiology community has started to achieve this goal. We believe the odyssey needs to continue (www.ohdsi.org), with many obstacles to be overcome, but with worthy rewards not just at the end of the journey but along the way as well.

Conflicts of interest The OMOP was funded by the Foundation for the National Institutes of Health (FNIH) through generous contributions from the following: Abbott, Amgen Inc., AstraZeneca, Bayer Healthcare Pharmaceuticals, Inc., Biogen Idec, Bristol–Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Janssen Research and Development, Lundbeck, Inc., Merck & Co., Inc., Novartis Pharmaceuticals Corporation, Pfizer Inc., Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi-aventis, Schering-Plough Corporation, and Takeda.

Patrick Ryan, Paul Stang and Martijn J. Schuemie are employed by and hold stock/stock options in Janssen Research & Development but have no conflicts of interest related to the content of this letter. J. Marc Overhage is employed by and holds stock/stock options in Siemens Health Services but has no conflicts of interest related to the content of this letter.

References

1. Hennessy S, Leonard CE. Comment on: "Desideratum for evidence-based epidemiology". Drug Saf. doi:10.1007/s40264-014-0252-x.
2. Madigan D, et al. A systematic statistical approach to evaluating evidence from observational studies. Annu Rev Stat Appl. 2014;1:11–39.
3. Ryan PB, et al. Defining a reference set to support methodological research in drug safety. Drug Saf. 2013;36(Suppl 1):S33–47.
4. Norén GN, Caster O, Juhlin K, Lindquist M. Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance. Drug Saf. 2014;37(9):655–9.
5. Murray RE, Ryan PB, Reisinger SJ. Design and validation of a data simulation model for longitudinal healthcare data. AMIA Annu Symp Proc. 2011;2011:1176–85.
6. Ryan PB, Schuemie MJ. Evaluating performance of risk identification methods through a large-scale simulation of observational data. Drug Saf. 2013;36(Suppl 1):171–80.
7. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005;58(4):323–37.
8. Corser W, et al. Concordance between comorbidity data from patient self-report interviews and medical record documentation. BMC Health Serv Res. 2008;8(1):85.
9. Tisnado DM, et al. What is the concordance between the medical record and patient self-report as data sources for ambulatory care? Med Care. 2006;44(2):132–40.
10. Weissman JS, et al. Comparing patient-reported hospital adverse events with medical record review: do patients know something that hospitals do not? Ann Intern Med. 2008;149(2):100–8.

11. Luck J, et al. How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. Am J Med. 2000;108(8):642–9.
12. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol. 2005;34(6):1370–6.
13. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. New York: Springer; 2009: p. 94–99.
14. Greenland S. Bias analysis. International encyclopedia of statistical science. Berlin: Springer; 2011: p. 145–148.
15. Stang PE, et al. Health outcomes of interest in observational data: issues in identifying definitions in the literature. Health Outcomes Res Med. 2012;3(1):e37–44.
16. Reich CG, Ryan PB, Schuemie MJ. Alternative outcome definitions and their effect on the performance of methods for observational outcome studies. Drug Saf. 2013;36(Suppl 1):S181–93.
17. Stang PE, et al. Variation in choice of study design: findings from the epidemiology design decision inventory and evaluation (EDDIE) survey. Drug Saf. 2013;36(Suppl 1):S15–25.