Editorial

Toward the Estimation of Unbiased Disease Prevalence Estimates Using Administrative Health Records


Data are information. And what we do with that information, how we process it, and interpret it can be complicated. It should not come as a surprise that these days there is a lot of talk about "big data" – about its promise, its potential, and its pitfalls. Big data (e.g., administrative, birth certificates, claims, electronic health records, registers) are growing in size, accessibility, and application. However, repurposing data from their original use to the research environment requires careful attention. Truthfully, whether we are talking about statistical analysis of small clinical datasets or supervised learning algorithms in big datasets, some of the same principles apply. No matter what, understanding where our data come from informs our design, our analysis, and most importantly, our interpretation.

There are 3 major sources of bias that determine whether inferences from a dataset are a close approximation of the truth: confounding, selection, and information. Confounding occurs when an association between 2 factors can be explained by an (often unmeasured) extraneous factor. Confounding often limits our ability to make truthful inferences about causality. Selection bias may occur when the choice of dataset limits the ability to generalize findings to the population affected by a disease. For example, using only drug claims data or hospitalization data to infer the prevalence of osteoarthritis (OA) may underestimate the condition because there may be individuals who may not need medication or have not been hospitalized in the time window evaluated. Information bias (often referred to as misclassification or measurement error) is also a threat to validity. Despite the potential problems of misclassification and measurement error, a recent systematic review found that fewer than 50% of studies from 12 high-impact journals in 2016 reported on this error, and only 7% used methods to assess or adjust for it[1]. Large samples alone cannot overcome systematic errors. In other words, infinitely large sample sizes will not necessarily mitigate these biases.

In this issue of *The Journal,* Slim, *et al*[2], use health administrative data from Quebec to estimate the prevalence of rheumatoid arthritis (RA) in 2010. Using about 20,000 participants aged 40 to 69 years old from the large prospective CARTaGENE cohort, the authors linked to the Régie de l'assurance maladie du Québec for provincial administrative health data. The choice of this dataset for the estimation of the prevalence of RA is appropriate and limits selection bias because Canada has universal healthcare and the administrative datasets that house the physician billing data are a close approximation of the true patterns of disease burden, or at the very least what physicians diagnose and document in the electronic health records. The authors demonstrate how the data one chooses can influence results and also highlight the importance of addressing misclassification. By increasing the observation period, cases can be identified that may be milder, untreated, or managed predominantly by primary care during shorter time windows. Using Swedish population-based registry data from 2001 to 2007, we found a comparable prevalence of RA in 2008, presented age-stratified as 0.19% (40–49 yrs), 0.43% (50–59 yrs), and 0.89% (60–69 yrs) on the basis of visits to inpatient or outpatient specialist care or entry in the Swedish Rheumatology Quality Register[3]. The results are intuitive – adding self-reported data increased the prevalence compared to using administrative data alone, and adjusting for the potential false positives reduced the prevalence.

Across all modeling approaches to estimate the prevalence of RA, the authors showed how increasing the observation period of the data influences the estimated prevalence. With all followup time ending on December 31, 2010, the prevalence of RA increased as the duration of followup time increased. The authors and others have demonstrated this in other settings including those for systemic lupus erythematosus (SLE) and OA[4,5,6,7,8]. The authors also acknowledged that there are some pitfalls in the use of longer observation periods and self-reported information on RA, with both methods increasing the risk of the overestimation of the true prevalence. To combat these risks and boost the reliability of the prevalence estimates, they included misclas-

See RA prevalence in Quebec, *page 1570*

sification error estimates and augmented self-reports of RA diagnosis with current use of disease-modifying anti-rheumatic drugs.

What this work adds to our dialogue is the importance of evaluating the likelihood of and accounting for potential misclassification. In the current setting, misclassification can happen 2 ways: (1) an individual is identified as having RA but does not actually have RA (a false positive), and (2) an individual is identified as not having RA but actually does (a false negative).

Remember that sensitivity is the probability of someone who has RA being identified by the algorithm as an RA case (true positive) and the specificity is the probability that an individual is correctly identified as not having RA (true negative). The complement of the latter, 1-specificity, is therefore the probability that an individual is falsely identified as a case [a false positive, i.e., Pr(Algorithm+|RA disease-)]. As the authors explain, the observed cases identified will always be some mix of true positives and false positives (i.e., a function of sensitivity and specificity). Hence, the authors use data on these variables informed by a published validation study and experts in the field, to adjust their estimates for this anticipated misclassification.

Misclassification and measurement errors are gaining more attention as sources of bias to be addressed in epidemiological research. Plotting the number of times these were mentioned by searching PubMed since 1995, we see that clearly more papers are considering this potential threat to validity (Figure 1). In the work by Slim, *et al*[2], the authors applied Bayesian latent class analysis, incorporating prior sensitivity and specificity of the ascertainment methods into the likelihood function and acknowledging the unknown truth (i.e., gold standard of confirmed RA). Electronic health records–based phenotyping using Bayesian latent class analysis has also recently been applied to type 2 diabetes mellitus and is also covered elsewhere in detail[9]. Additional approaches, including non-Bayesian methods, and considerations for quantitative bias analysis are discussed by Lash and colleagues, including consideration of when bias analysis may be more helpful than necessary[10].

Confounding, selection bias, and misclassification are critical threats to the validity and generalizability of our work, and the size and availability of large datasets is only going to increase. In an extreme example, our group showed that death certificate data may underestimate the burden of SLE in Sweden, with about 59% of decedents with SLE lacked mention of SLE on their death certificates[11]. We determined that certain characteristics lead to missingness (older age and having a cancer diagnosis), and that there are situations where the extent and direction of the misclassification may be impossible to quantify. However, with administrative datasets such as the one used in this study, the level of misclassification is often not as stark as the limited death certificate data. We are getting more comfortable with the notion of confounding and the myriad strategies to tackle this potential source of bias. Measurement error and misclassification exist, whether we acknowledge them or not, and may not always simply lead to conservative estimates by biasing toward the null or diluting estimates of the truth. Thus, it is important to understand the provenance of the data and how that informs our interpretation.

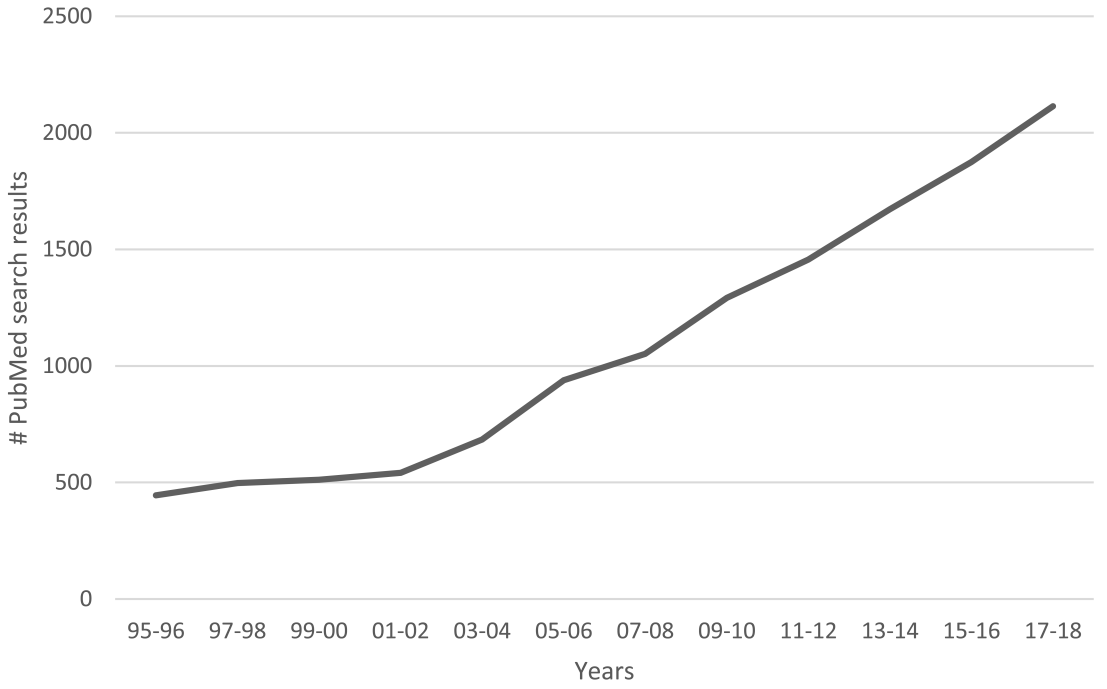

Figure 1. PubMed search for papers that mentioned misclassification and measurement errors as sources of bias.

TITILOLA FALASINNU, PhD,
Postdoctoral Fellow,
Division of Epidemiology,
Department of Health Research and Policy,
Stanford School of Medicine;
JULIA F. SIMARD, ScD,
Assistant Professor,
Division of Epidemiology,
Department of Health Research and Policy,
and Division of Immunology and Rheumatology,
Department of Medicine, Stanford School of Medicine,
Stanford, California, USA.

Address correspondence to J.F. Simard, Assistant Professor, Division of Epidemiology, Department of Health Research and Policy, Stanford School of Medicine, HRP Redwood Building, Room T152, 259 Campus Drive, Stanford, California 94305-5405, USA.
E-mail: jsimard@stanford.edu

REFERENCES

1. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. J Clin Epidemiol 2018;98:89-97.
2. Slim ZF, Soares de Moura C, Bernatsky S, Rahme E. Identifying rheumatoid arthritis cases within the Quebec health administrative database. J Rheumatol 2019;46:1570-6.
3. Neovius M, Simard JF, Askling J. Nationwide prevalence of rheumatoid arthritis and penetration of disease-modifying drugs in Sweden. Ann Rheum Dis 2011;70:624-9.
4. Ng R, Bernatsky S, Rahme E. Observation period effects on estimation of systemic lupus erythematosus incidence and prevalence in Quebec. J Rheumatol 2013;40:1334-6.
5. Nightingale AL, Farmer RD, de Vries CS. Systemic lupus erythematosus prevalence in the UK: methodological issues when using the General Practice Research Database to estimate frequency of chronic relapsing-remitting disease. Pharmacoepidemiol Drug Saf 2007;16:144-51.
6. Wiréhn A-BE, Karlsson HM, Carstensen JM. Estimating disease prevalence using a population-based administrative healthcare database. Scand J Public Health 2007;35:424-31.
7. Powell KE, Diseker RA, Presley RJ, Tolsma D, Harris S, Mertz KJ, et al. Administrative data as a tool for arthritis surveillance: estimating prevalence and utilization of services. J Public Health Manag Pract 2019;9:291-8.
8. Kopec JA, Rahman MM, Berthelot J-M, Le Petit C, Aghajanian J, Sayre EC, et al. Descriptive epidemiology of osteoarthritis in British Columbia, Canada. J Rheumatol 2007;34:386-93.
9. Hubbard RA, Huang J, Harton J, Oganisian A, Choi G, Utidjian L, et al. A Bayesian latent class approach for EHR-based phenotyping. Stat Med 2019;38:74-87.
10. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol 2014;43:1969-85.
11. Falasinnu T, Rossides M, Chaichian Y, Simard JF. Do death certificates underestimate the burden of rare diseases? The example of systemic lupus erythematosus mortality, Sweden, 2001-2013. Public Health Rep 2018;133:481-8.

J Rheumatol 2019;46:1549–51; doi:10.3899/jrheum.190484