

# Routinely collected data: the importance of high-quality diagnostic coding to research

Stuart G. Nicholls PhD, Sinéad M. Langan MSc PhD, Eric I. Benchimol MD PhD

■ Cite as: *CMAJ* 2017 August 21;189:E1054-5. doi: 10.1503/cmaj.170807

See related article at [www.cmajopen.ca/content/5/3/E617](http://www.cmajopen.ca/content/5/3/E617)

**R**outinely collected health data are data collected for purposes other than research or without specific a priori research questions developed before collection.<sup>1</sup> Examples include clinical information from electronic health records, health administrative data, disease registries and epidemiologic surveillance systems.<sup>1</sup> Health data of this type are used widely for clinical, pharmacoepidemiologic and health services research.<sup>2</sup> However, the quality of these data remain in question. Tang and colleagues consider this issue in a linked qualitative research article published in *CMAJ Open*.<sup>3</sup> The authors examine barriers to coding of high-quality administrative data and highlight several factors that may lead to a lack of specificity of diagnoses or data inconsistencies between the medical chart and administrative data. In particular, they point to the role played by the quality of physician notes.

Studies mapping the coding process for diagnoses have identified a range of sources of error along both the patient diagnostic trajectory, as well as errors that occur during the administrative process.<sup>4</sup> Tang and colleagues' research contributes to the growing awareness of the need to develop and validate approaches to accurately identifying patients according to exposure or outcome when using health administrative data for research,<sup>5,6</sup> and raises important questions about how best to address inaccuracies. Specifically, it broaches the question of who has responsibility to improve accuracy. The short answer is that every link in the chain has a role because each step may have flaws, from the physician and their notes (the focus of the study by Tang and colleagues),<sup>3</sup> to the hospital and its structure, to the use of International Classification of Disease (ICD) codes themselves, as well as coders and their training.

Sources of error often act in concert: physicians recognize new diagnoses, yet the adoption of new diagnoses within the ICD-10 system is slow, potentially leading to confusion between the physicians' notes and coders who were trained only to identify diseases in the coding system used by the administration. For example, eosinophilic esophagitis was recognized as a clinical entity distinct from reflux esophagitis in 1995,<sup>7</sup> but the ICD-10 code adopted in 2015 (K20.0) is still not in use by most agencies that collect data. Furthermore, the physician billing systems of

## KEY POINTS

- Errors in diagnostic coding may occur along the diagnostic or administrative pathways.
- There is an ongoing need for validation work to inform research using health administrative data.
- The quality of reporting is key to the evaluation of research using health administrative data, and authors should conform to standards of best practice and use appropriate reporting guidelines.

many Canadian provinces continue to use an antiquated and simplified version of ICD-9. This can lead to tremendous confusion and inaccuracies in research that relies upon these data. Health care administrators and policy-makers should be encouraged to adopt the latest diagnostic systems and to train coders and physicians appropriately in their use.

In using coded routinely collected data, validation of coding accuracy is key. This is important for both the methodologic evaluation of articles ("Can the results be trusted?")<sup>8</sup> and for replication ("Do initial findings hold true in other contexts?"). For example, the accuracy of reports about the prevalence of a disease, or the burden of a disease on the health system, will rely on the accuracy of identification algorithms based on diagnostic codes.<sup>9</sup> Given the potential for misclassification bias, it is essential that researchers report the results of validation work.<sup>5</sup> Despite this need, a recent review of published studies<sup>10</sup> showed that the reporting of research using routinely collected health data often inadequately provided details of validation work or information about coding or classification.

Furthermore, those who use health administrative data for research, health care administration or quality improvement have their own obligations to use the data responsibly and to report accurately how data were used.<sup>2</sup> For studies using routinely collected data, the REporting of studies Conducted using Observational Routinely-collected Data (RECORD) statement<sup>1</sup> sets out standards of reporting that researchers should adhere to. This statement includes 13 items specific to studies using

routinely collected data and reflects important study components such as methods of selecting the study population; details of any validation of codes or algorithms; linkage of data sources; and a list of codes that are used to classify exposures, outcomes and confounders. Adherence to these guidelines will allow the consumer of research to accurately interpret the results, a bare minimum requirement for medical research.

Although the study by Tang and colleagues<sup>3</sup> evaluates barriers to adequate coding at present, there are many unanswered questions. The authors have suggested improved training of physicians to understand the importance of accurate documentation. However, such interventions have not yet been designed or evaluated. It may be better to target this type of intervention at medical students still in training so that the importance of coding is understood by trainees early in their careers. In addition, the linked article did not address the changes that might result from widespread implementation of electronic health records in hospitals. The mix of clinical documentation and structured coding in electronic health records may result in improved interpretation by professional coders. The increasing availability of natural language processing techniques and machine learning may lead to automated interpretation of physician documentation, more accurate coding tools and automated feedback to physicians who provide contradictory diagnoses in the medical record.

Although the use of artificial intelligence in health data research is expected to blossom in the future, current research is still largely curated and conducted by people. Therefore, we must strive to improve the quality of coding (through training of

physicians and coders, and implementation of the latest coding systems), the quality of research using these data (through validation and advanced methodologies), and the quality of reporting of the research (through the use of reporting guidelines to improve the transparency and reproducibility of research).

## References

1. Benchimol EI, Smeeth L, Guttman A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 2015;12:e1001885.
2. Nicholls SG, Langan SM, Benchimol EI. Reporting and transparency in big data: the nexus of ethics and methodology. In: Mittelstadt BD, Floridi L, editors. *The ethics of biomedical big data*. Switzerland: Springer; 2016:339-65.
3. Tang K, Lucyk K, Quan H. Coder perspectives on physician-related barriers to coding high quality administrative data: a qualitative study. *CMAJ Open* 2017; 5:E617-22.
4. O'Malley KJ, Cook KF, Price MD, et al. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005;40:1620-39.
5. Benchimol EI, Manuel DG, To T, et al. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;64:821-9.
6. De Coster C, Quan H, Finlayson A, et al. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Serv Res* 2006;6:77.
7. Kelly KJ, Lazenby AJ, Rowe PC, et al. Eosinophilic esophagitis attributed to gastroesophageal reflux: improvement with an amino acid-based formula. *Gastroenterology* 1995;109:1503-12.
8. Moher D, Simeria I, Schulz KF, et al. Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research. *BMC Med* 2008;6:13.
9. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ* 2010;341:c4226.
10. Hemkens LG, Benchimol EI, Langan SM, et al. The reporting of studies using routinely collected health data was often insufficient. *J Clin Epidemiol* 2016;79:104-11.

**Competing interests:** Sinéad Langan has received a grant in the form of a Wellcome Senior Clinical Fellowship in Science. Eric Benchimol is supported by a New Investigator Award from the Canadian Institutes of Health Research, Canadian Association of Gastroenterology, and Crohn's and Colitis Canada. He is also supported by the Career Enhancement Program of the Canadian Child Health Clinician Scientist Program. No other competing interests were declared.

**Affiliations:** Children's Hospital of Eastern Ontario (CHEO) Research Institute (Nicholls, Benchimol); School of Epidemiology and Public Health, Faculty of Medicine (Nicholls, Benchimol), University of Ottawa, Ottawa, Ont.; London School of Hygiene and Tropical Medicine (Langan), London, UK; Department of Pediatrics (Benchimol), University of Ottawa, Ottawa, Ont.

**Contributors:** All of the authors contributed to the conception of the work, drafted and

revised the manuscript for important intellectual content, provided final approval of the version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

This article was solicited and has not been peer reviewed.

**Correspondence to:** Stuart Nicholls, SNicholls@cheo.on.ca