Guiding the reporting of studies that use routinely collected health data

Hude Quan MD PhD, Tyler Williamson PhD

See also www.cmaj.ca/lookup/doi/10.1503/cmaj.150653

In the last decade, there has been an explosion of digital information, and health information is no exception. Aided by the development of technology, vast amounts of data are being generated in health care, including abstracted clinical data from hospital discharge summaries, physician billing data and patient health information from electronic medical records (EMRs). These data are rich in clinical information on diagnoses, procedures, medications, laboratory results and imaging information. Although such data are collected for health care management or clinical care, they have enormous potential for use in disease surveillance and answering clinical questions through research-based hypotheses. However, routinely collected data are often limited because data structures and medical terminologies are not standardized across jurisdictions and countries. For example, information can be entered as structured and coded data (e.g., using International Classification of Disease [ICD] codes) or as unstructured text (e.g., in EMRs).

The challenges posed by analyzing data with different terminologies, quality and collection methods have led to the development of a new reporting statement and checklist: the RECORD (Reporting of Studies Conducted using Observational Routinely Collected Health Data).[1] Based on the well-known STROBE statement (Strengthening the Reporting of Observational Studies in Epidemiology), the RECORD statement expands on specific items in STROBE that are applicable to routinely collected health data. For example, it expands the requirement of reporting all variables by encouraging authors to report a complete list of the codes and algorithms used throughout the analysis (e.g., in classifying the outcome and exposure). In some scenarios, data linkages are based on a unique identifier and are deterministic; however, in the absence of a unique patient identifier across data resources, a probability linkage method is frequently adopted. The completeness and accuracy of the probability linkage may af-

fect selection bias in the analysis; therefore, the RECORD statement suggests that the method and rate of linkage should be reported to allow readers to judge the risk of selection bias.

The importance of the RECORD statement is proportional to the value of routinely collected data for generating new knowledge and advancing health science, which is increasingly being recognized by many countries worldwide. For instance, data linkage, analysis and reporting centres have been established in Canada, Australia and the United Kingdom.[2] They are pioneering the collection and use of big health data and are showing their power and impact. Great strides have been made, particularly in four major categories: disease surveillance, health services and outcome research, monitoring population health and health system performance, and precision medicine.

In Canada, two major groups are engaged in chronic disease surveillance: the Canadian Chronic Disease Surveillance System and the Canadian Primary Care Sentinel Surveillance Network. The former uses ICD-coded hospital administrative data and physician claims data,[3] and the latter extracts primary care EMR data.[4]

The Interdisciplinary Chronic Disease Collaboration in Alberta has been a leader in health services and outcome research by establishing population-based cohorts to study hypertension, diabetes, chronic kidney disease and vascular disease using routinely collected administrative health data. The group has analyzed these data to

---

KEY POINTS

- Routinely collected health data have been widely analyzed for many purposes, particularly surveillance and research.
- There are inconsistencies in the reporting of results arising from the analysis of routinely collected data.
- Studies conducted using routinely collected data should be reported following the new RECORD (Reporting of Studies Conducted using Observational Routinely Collected Health Data) statement.

All editorial matter in *CMAJ* represents the opinions of the authors and not necessarily those of the Canadian Medical Association.

explore interactions among chronic diseases, such as the effect of diabetes on myocardial infarction and all-cause mortality.[5]

Traditional data-collection methods (e.g., surveys and chart reviews) have become too costly to be used frequently to monitor population health and health system performance. The Canadian Institute for Health Information has maximized efficiency by centralizing huge amounts of national, routinely collected data from which numerous reports on population health status and health system performance in Canada have been generated.[6]

An innovative and promising use of routinely collected health data is in precision medicine; that is, precise prognostication for an individual patient based on the treatment received and his or her sociodemographic, genetic, environmental and clinical characteristics.[7] Public health organizations aim to create precise forecasts of infectious disease epidemics, monitor preventive intervention programs, and measure and address health disparities. Such activities require large volumes of high-quality data and corresponding analytics.

Maximizing the health benefits of routinely collected data requires the development of methods to avoid "garbage in and garbage out." New ways of harmonizing, linking and structuring data will help to generate new knowledge. In addition, systematic strategies to assess and improve data quality, such as automated programs to detect coding errors, should be developed and implemented. Machine learning, which uses statistical and computer science methods to identify patterns and characteristics within empirical data, can lead to new knowledge and improve decision-making, all of which will result ultimately in better health care delivery with greater operational efficiency.

The RECORD statement makes at least three important contributions to aid users of routinely collected health data: (a) a method for assessing the quality of the reporting, (b) a framework for development of a research protocol and (c) a guide for best practices in manuscript writing. Authors of studies who use and analyze routinely collected data should apply the RECORD statement when reporting their specific data analyses. The consistent and regular use of the statement should improve quality, clarity and replicability of their work. The statement is internationally applicable and should enhance the quality of reporting in this area. It provides an important framework for the reporting of results generated from the analysis of routinely collected data.

References

1. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015;12:e1001885.
2. Lucyk K, Lu M, Sajobi T, et al. Administrative health data in Canada: lessons from history. *BMC Med Inform Decis Mak* 2015; 15:69.
3. Robitaille C, Dai S, Waters C, et al. Diagnosed hypertension in Canada: incidence, prevalence and associated mortality. *CMAJ* 2012;184:E49-56.
4. Williamson T, Green ME, Birtwhistle R, et al. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. *Ann Fam Med* 2014;12:367-72.
5. Tonelli M, Muntner P, Lloyd A, et al. Risk of coronary events in people with chronic kidney disease compared with those with diabetes: a population-level cohort study. *Lancet* 2012;380:807-14.
6. *Trends in income-related health inequalities in Canada.* Ottawa: Canadian Institute for Health Information; 2015. Available: https://secure.cihi.ca/free_products/trends_in_income_related_inequalities_in_canada_2015_en.pdf (accessed 2016 Jan. 20).
7. Ashley EA. The precision medicine initiative: a new national effort. *JAMA* 2015;313:2119-20.

Affiliation: Department of Community Health Sciences, University of Calgary, Calgary, Alta.