Routinely collected data for population-based outcomes research

Blayne Welk, MD, MSc
Department of Surgery (Urology) and Epidemiology & Biostatistics, Western University, London, ON, Canada

Cite as: *Can Urol Assoc J* 2019 July 23; Epub ahead of print.
http://dx.doi.org/10.5489/cuaj.6158

Published online July 23, 2019

***

Introduction

Routinely collected data (or administrative data), is a source of data for many studies that assess a variety of questions such as epidemiological trends over time to clinically relevant associations between risk factors and disease. This data comes from databases that record information for a purpose other than medical research, such as for hospital or physician reimbursement.

There are several strengths of routinely collected data studies:

1. Low study costs
2. Rapid study completion
3. Good for estimating incidence/prevalence in a population
4. Often have large sample sizes and significant statistical power
5. Better generalizability to the real world
6. Prolonged retrospective study periods are possible
7. Longitudinal followup across providers and regions may be possible
8. Improved feasibility for studying rare populations, exposures and outcomes
9. Can study outcomes or exposures that would be unethical in a prospective study
10. Well suited for measuring geographical variation

There are also potential limitations that must be considered when conducting or reading a routinely collected data study:

1. The validity and reliability of the data elements may be poor
2. Often not all clinically relevant variables are present
3. Results may not be hypothesis driven and could represent a spurious association or demonstrate a statistically significant result that is not clinically relevant.
4. Data collection methods or coding practices may change over time, and this may not be evident to the researcher.

Epidemiological considerations

Routinely collected data is usually used to either describe a something (for example incidence of a disease, changes in treatment over time, or resource utilization) or to perform an observational study. Observational studies have potential biases associated with them, of which a few are particularly relevant to those that use routinely collected data.

1. Selection bias occurs when a study population is not a random sample from the target population that you wish to generalize your results to. For example, most randomized controlled trials have strict inclusion/exclusion criteria, however physicians use the interventions studied in those trials on patients who would not have been eligible for randomized trial with the assumption that the results will be similar.
2. Information bias occurs when the variable is not measured accurately. This lead to either misclassification, or measurement errors. While prospective studies can explicitly define a method of measurement that maximizes accuracy (for example taking 3 blood pressure readings, 3 minutes apart after the patient has rested in the seated position for 2 mins), this is usually not the cause with routinely collected data variables. This is because the administrative data elements are not created or recorded for the purposes of research, and often indicator variables are used to represent a clinical condition (for example in a clinical study pathology data would be used to determine if a patient had prostate cancer, whereas in an administrative data study, a physician code for the performance of a radical prostatectomy might be used as a marker for prostate cancer). If misclassification or measurement error is random, it biases the results towards a null association, as confidence intervals widen due to more "noise" in the data. If it is not random, this can significantly affect the results and lead to completely mistaken conclusions.[1]

> *How well do the key variables (such as the codes used to identify the population, primary exposure and primary outcome) represent what the research is actually interested in?*
>
> Consider how common the condition is, how likely is that the coding element would be recorded, how likely the coding element could be confused for another condition or procedure, what measures the database has to ensure correct codes are entered, and what the motivations are of the people submitting the coding elements. Ideally these key variables such as the primary outcome should have known measurement characteristics (such as a positive predictive value) so that you can judge how well that code represents what it is meant to represent. This has traditionally been poorly done, [2-4] and when it is done this elevates administrative data studies to a higher level.

3. Confounding occurs when with the relationship between an exposure and outcome is distorted by another variable, which acts as a confounder. Known confounders can be controlled for, however unknown or unmeasured confounders can only be properly controlled for with randomization, which is not possible with retrospective administrative data studies. Propensity scores and instrumental variables can help address confounder, but does not eliminate the risk of residual confounding.[5]

Transparent reporting of a routinely collected data study

Most physicians are aware of reporting standards for randomized clinical trials (CONsolidated Standards Of Reporting Trials, CONSORT). This guideline has improved the quality of clinical trial reporting. An analogous reporting guideline is available for routinely collected data studies (RECORD: REporting of studies Conducted using Observational Routinely-collected health Data).[6] Similar to this reporting guideline, others have proposed criteria to evaluate the quality of administrative database studies[7]:

Methodological principle
Study design clearly described
Administrative database comparative study
Administrative database case–control study
Administrative database case series
Why database was created clearly stated
Description of database's inclusion/exclusion criteria
Description of methods for reducing bias in database
Codes and search algorithms reported
Rationale for coding algorithm reported
Code accuracy reported
Code validity reported
Clinical significance assessed
Is the period of data consistent with the outcome data?
Statement regarding whether data stems from single or multiple hospital admissions
Statement regarding whether data stems from single or multiple procedures
Accounting for clustering

Adapted from Hashim et al, Evidence-based spine-care journal, Oct 2014.

Examples and brief overview of routinely collected data sources

	Description	**Major data elements**
Surveillance, Epidemiology, and End Results Program (SEER)[8]	United States cancer registry which includes approximately 35% of the US population. Data are representative of the US population and are drawn from 12 state registries, 4 metropolitan multicounty areas, and 3 indigenous registries	Patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and followup for survival
Medicare/Medicaid[9]	National records of reimbursement related to subsidized care provided to US citizens >65 years of age (Medicare), or low income adults, those with a physical disability, and children (Medicaid)	Part A covers non-physician inpatient care, Part B covers physician services, and Part D includes optional drug coverage. Demographic and geographic information, diagnosis (ICD code) and procedures (CPT or HCPC codes) and national drug codes are included in each respective part.
National Inpatient Sample (NIS)	National representative sample of discharges (20%) of children and adults from all community	Discharge abstracts include ICD codes for admission and discharge diagnoses, demographics,

	hospitals (includes those with both Medicare/Medicaid, private insurance, and no insurance)	hospital characteristics, payment source, length of stay, severity and comorbidity measures.
American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP)	Voluntary hospital level program that compares risk-adjusted outcomes after surgical procedures. Over 650 hospitals (primarily from the United States) are participating in order to compare their post-surgical complications to national averages.	Demographics, operative procedure (CPT code), selected risk factors (such as diabetes, smoking, medical comorbidities), preoperative laboratory values, length of stay, and specific complications that occur within 30 days of the initial OR (such as unplanned reoperation, stroke, bleeding, UTI, and wound infection)

Conclusions

Electronic data is a driving force in our society. It has an annual compound growth of 60%, and in 2020 it is estimated there will be 35 zettabytes of electronic data.[10] In healthcare, information technology plays a key role in all aspects of practice, from medical records to medication prescribing to communication. This wealth of readily available electronic information will likely continue to drive medical research using routinely collected data. An *a priori* hypothesis and analytical plan, valid data elements, appropriate statistical techniques, a careful assessment of bias, and high-quality reporting will hopefully continue to improve the quality and impact of these studies in urology. Despite the limitations of observation studies, they often produce results similar to randomized controlled trials.[11] Other well written reviews specific to urologists have been published[12,13] and are worth reviewing for those interested in administrative data research.

References

1. Höfler M: The effect of misclassification on the estimation of association: a review. Int J Methods Psychiatr Res 2005; 14: 92–101.
2. 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 2010: 1–9.
3. van Walraven C, Bennett C and Forster AJ: Administrative database research infrequently used validated diagnostic or procedural codes. J Clin Epidemiol 2011; 64: 1054–1059.
4. Welk B and Kwong J: A review of routinely collected data studies in urology: Methodological considerations, reporting quality, and future directions. Can Urol Assoc J 2017; 11: 136–6.
5. Normand SLT, Sykora K, Li P, et al: Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. BMJ: British Medical Journal 2005; 330: 1021.
6. 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.
7. Hashimoto RE, Brodt ED, Skelly AC, et al: Administrative database studies: goldmine or goose chase? Evid Based Spine Care J 2014; 5: 74–76.
8. Engels EA, Pfeiffer RM, Ricker W, et al: Use of Surveillance, Epidemiology, and End Results-Medicare Data to Conduct Case-Control Studies of Cancer Among the US Elderly. American Journal of Epidemiology 2011; 174: 860–870.
9. Mues KE, Liede A, Liu J, et al: Use of the Medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US. Clin Epidemiol 2017; 9: 267–277.
10. Data Universe Explosion & the Growth of Big Data | CSC. 2016: 1–3. Available at: http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode, accessed May 24, 2016.
11. Anglemyer A, Horvath HT and Bero L: Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. (Edited byL Bero). Chichester, UK: John Wiley & Sons, Ltd; 1996:1–46.
12. Schlomer BJ and Copp HL: Secondary Data Analysis of Large Data Sets in Urology: Successes and Errors to Avoid. J. Urol. 2014; 191: 587–596.
13. Cole AP, Friedlander DF and Trinh Q-D: Secondary data sources for health services research in urologic oncology. Urol. Oncol. 2018; 36: 165–173.