Challenges to using big data in health services research

Hosung Kang, Shannon Sibbald

ABSTRACT

Given the shift in current healthcare trends toward digitization of storing information, there has been an increase in the number of studies using administrative databases. These databases provide a powerful tool to conduct research on outcomes, health services, and epidemiology. However, these databases have limitations and biases that should be considered. Given the sensitive information regarding patients' health in the database, security clearances must be granted before data is accessed. Furthermore, algorithms to link the different variables to create a cohort of people with specific disease are imperfect and may not yield an accurate representation. Due to a large volume of records, a statistically significant finding may be observed, but may provide insignificant clinical results. Despite the current limitations, administrative databases provide powerful data that researchers can use to identify gaps in performance to improve the healthcare system.

INTRODUCTION

Healthcare has shifted considerably toward digitization in the use of monitoring and storing information.¹ This has increased the ability to use administrative health databases to conduct health services and clinical research.² Administrative databases record information that is routinely collected across the health system (eg in family health clinics, hospitals, pharmacy visits, and laboratory tests). There are several organizations that collect information and conduct research using these databases, including the Institute for Clinical Evaluative Sciences (ICES), Canadian Institute for Health Information (CIHI), and Research Data Centres (RDC). Researchers at ICES, for example, continuously publish articles regarding wait times, health outcomes, health services and other articles that are of importance to the general public regarding the current healthcare system.³ Large databases can provide a powerful tool to conduct research into epidemiology, effectiveness, and safety outcomes.^{4,5} However, there are several limitations and challenges in both accessing the databases and in utilizing them effectively.

ETHICS, PRIVACY, AND CONFIDENTIALITY CONSIDERATIONS

Administrative databases contain sensitive information about a patient's health, diagnoses, demographics, and where they reside. All this information can be used to deduce other factors such as socioeconomic status.⁶ Information stored in administrative databases is inherently protected in Canadian laws governing rights and privacy. There is legislation safeguarding the storage and access to this information, such as Ontario's Freedom of Information and Protection of Privacy Act (FIPPA) and Personal Health Information Protection Act (PHIPA).^{7,8} Since privacy and confidentiality are of

utmost importance, thorough clearance checks are required before individuals or organizations can be granted access to the databases, which can cause delays in accessing data.⁶ To safeguard anonymity, data is often transformed, excluded, and/or encrypted before being stored in the databases.⁶ This causes a time lag between when the data is recorded and when it is transferred to the databases. Thus, conducting research in 'real time' may be very challenging. Consequently, studies using administrative databases are limited to retrospective studies. Researchers and clinicians employing a health intervention may not know the results until months or years after the intervention has been completed.

RULES AND REGULATIONS AS BARRIERS

RDC, CIHI, and ICES provide guidelines on navigating the steps required to access databases. For example, CIHI allows graduate students working on a project using administrative databases to request access through the Graduate Student Data Access Program (GSDAP).⁹ However, additional barriers exist. For example, if the supervisor has funding to conduct the research, students cannot file an application through the GSDAP.⁹ Independent researchers affiliated with a university require funding to access databases, which often poses a financial barrier to such research. The minimum funding required to conduct a rudimentary study can reach tens of thousands of dollars, with the average funding required to conduct more thorough research being significantly higher.¹¹ Additionally, students wanting to access these health databases to conduct research in fulfilment of their independent project, thesis, or dissertation may require a faculty supervisor affiliated with the organization providing access to the databases. This poses another challenge because students are required to find a supervisor that not only matches their research interests, but also someone who is a member of the desired organization. For example, ICES has a student program only available to students who have acquired supervision from a faculty member affiliated with ICES. RDC and CIHI, on the other hand, do not require a faculty member supervision affiliated with these organizations before access to the databases are granted, thereby providing easier access.^{9,10} Publicly funded faculty members and students can submit a proposal to ICES, which can reduce costs but does not eliminate them entirely.

RELIABILITY OF THE DATA AND CODE ACCURACY

Administrative databases are typically intended for financial and administrative management rather than for research purposes. In other words, they were created to monitor utilization, determine the amount of resource consumption, or to determine the required capacity for a service. Although administrative data contains a large amount of information that can determine hospital performance and identify where the current healthcare system is lacking, the

degree of detail and accuracy of the data within the databases may vary for many reasons.¹² The quality of information in these systems depends on incentives of data reporting such as financial and systematic support.¹³ This has meant that expensive medical procedures are documented more thoroughly than less costly health services, such as ambulatory services.¹⁴ One of the most common types of error in these databases is absent values, which occurs when required information is missing or not inputted. This shows that administrative data may be incomplete and may not be the most accurate source of information.

One of the core elements of the administrative data system is the International Statistical Classification of Diseases and Related Health Problems (ICD; the most current version is ICD-11), which is a medical classification list by the World Health Organization (WHO).^{1,2,13} Using various algorithms to link specific ICD codes can provide a cohort of patients with the disease of interest, such as chronic obstructive pulmonary disease (COPD).¹⁵ For example, information such as cumulative patient profile, ICD codes, and prescriptions can be linked to identify individuals with a certain type of disease.¹⁶ However, there is a risk of bias, as the algorithms do not provide the highest level of accuracy. Fluctuations in the level of accuracy can be induced by including or excluding different variables.^{17,18} For example, Cooke et al were able to improve the accuracy of an algorithm that identifies patients with COPD by including albuterol, ipratropium, and age as additional variables.¹⁷

One indicator of the accuracy of an algorithm is the positive predictive value (PPV), which refers to the probability of presence of disease given a positive result.^{19,20} A study by Brenner and Gefeller have shown that PPV changes depending on the prevalence of the disease.¹⁹ By using a simulation study, they showed that PPV dropped from 95% to 60% when the disease prevalence decreased from 90% to 20% even though the algorithm did not change. Therefore, the algorithmically-predicted probability of a patient having the disease decreases when disease prevalence decreases.^{1,19,20} This is an important issue that needs to be considered when conducting research using administrative databases because most validation studies use samples consisting entirely of people with the appropriate ICD code instead of the population. A mixture of people with and without the ICD code will have lower disease prevalence than samples consisting entirely of people with the ICD code and will provide more accurate representation.¹

CLINICAL SIGNIFICANCE

One of the advantages of administrative database research is the large volume of data. However, because very small variations between groups in large databases result in statistically significant differences, caution should be taken in interpreting the results as clinically significant.¹ Absolute and relative values should be used instead of the P-value to gauge the importance of clinical differences. Since confidence intervals (CIs) are generated around the absolute or relative difference between populations, the use of CIs could help distinguish between clinical and statistical difference.

CONCLUSION

Current healthcare technology is shifting towards digitization of storing information, facilitating a subsequent increase in studies using administrative databases. Databases provide a powerful tool to conduct studies on outcomes, health services, and epidemiology. However, these databases have limitations and biases that should be considered when embarking on research. Given the presence of sensitive information in the database, access can be quite difficult and there are often strict policies in place as safeguard measures, which poses as an additional barrier. Furthermore, algorithms to link variables as means to create a cohort with a specific disease may not yield the most accurate representation. Due to a large volume of records, statistically significant results may be observed but may provide insignificant clinical results. Although there are drawbacks to using administrative databases, they still provide real-world data for researchers to use to improve the current healthcare system.

REFERENCES

1. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. *J Clin Epidemiol*. 2012 Feb 1;65(2):126–31.
2. Hashimoto RE, Brodt ED, Skelly AC, *et al*. Administrative database studies: goldmine or goose chase? *Evid Based Spine Care J*. 2014 Oct;5(2):74–6.
3. Burton KR, Park AL, Fralick M, *et al*. Risk of early onset breast cancer among women exposed to thoracic computed tomography in pregnancy or early postpartum. *J Thromb Haemost*. 2018 Feb 16;16(5):876–85.
4. McIsaac DI, Wong CA, Bryson GL, *et al*. Association of polypharmacy with survival, complications, and healthcare resource use after elective noncardiac surgery: a population-based cohort study. *Anesthesiology*. 2018 Jun;128(6):1140–50. <https://doi.org/10.1097/ALN.00000000000002124>
5. Cookson R, Mondor L, Asaria M, *et al*. Primary care and health inequality: Difference-in-difference study comparing England and Ontario. *PLOS ONE*. 2017 Nov 28;12(11):e0188560.
6. Krishna R, Kelleher K, Stahlberg E. Patient confidentiality in the research use of clinical medical databases. *Am J Public Health*. 2007 Apr;97(4):654–8.
7. Government of Ontario. Freedom of Information and Protection of Privacy Act. R.S.O. 1990, c. F.31. 2018 Apr 30 [cited 2018 Mar 9]. Available from: <https://www.ontario.ca/laws/statute/90f31>
8. Government of Ontario. Personal Health Information Protection Act, S.O. 2004, c. 3, Sched. A. 2018 May 7 [cited 2018 Mar 9] Available from: <https://www.ontario.ca/laws/statute/04p03>
9. CIHI. Graduate Student Data Access Program [Internet]. 2017 Feb 21 [cited 2018 Mar 9]. Available from: <https://www.cihi.ca/en/graduate-student-data-access-program>
10. Statcan. The Research Data Centres Program [Internet]. 2018 April 6 [cited 2018 July 31]. Available from: <https://www.statcan.gc.ca/eng/rdc/masters>
11. ICES [Internet]. Toronto (ON). Institute for Clinical Evaluative Sciences. [cited 2018 March 14]. Available from: <https://www.ices.on.ca/Research/Grants>
12. Hertzner NR. Reasons why data from the Nationwide Inpatient Sample can be misleading for carotid endarterectomy and carotid stenting. *Semin Vasc Surg*. 2012 Mar;25(1):13–7.
13. Freitas JA, Silva-Costa T, Marques B, *et al*. Implications of data quality problems within hospital administrative databases. In: XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010. Springer, Berlin, Heidelberg; 2010. p. 823–6. (IFMBE Proceedings).
14. Romano PS, Hussey P, Ritley D. Selecting quality and resource use measures: a decision guide for community quality collaboratives. Question 2: What are the

strengths and weaknesses of using administrative data, medical record data, and hybrid data? AHRQ Publication No. 09(10)-0073. Rockville, MD: Agency for Healthcare Research and Quality; 2010 May.

15. Gershon AS, Wang C, Wilton AS, *et al.* Trends in chronic obstructive pulmonary disease prevalence, incidence, and mortality in Ontario, Canada, 1996 to 2007: a population-based study. *Arch Intern Med.* 2010 Mar 22;170(6):560–5.
16. Lee TM, Tu K, Wing LL, *et al.* Identifying individuals with physician-diagnosed chronic obstructive pulmonary disease in primary care electronic medical records: a retrospective chart abstraction study. *NPJ Prim Care Respir Med.* 2017 May 15;27.
17. Cooke CR, Joo MJ, Anderson SM, *et al.* The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. *BMC Health Serv Res.* 2011 Feb 16;11:37.
18. Kadhim-Saleh A, Green M, Williamson T, *et al.* Validation of the diagnostic algorithms for 5 chronic conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a Kingston Practice-based Research Network (PBRN) report. *J Am Board Fam Med.* 2013 Apr;26(2):159–67.
19. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med.* 1997 May 15;16(9):981–91.
20. Parikh R, Mathai A, Parikh S, *et al.* Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol.* 2008 Feb;56(1):45–50.