Secondary Data Analysis of Epidemiology in Asia

Yoshitaka Murakami

Professor, Department of Medical Statistics, Toho University, Tokyo, Japan

Received July 22, 2014; accepted July 23, 2014; released online August 23, 2014

Copyright © 2014 Yoshitaka Murakami. This is an open access article distributed under the terms of Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Key words: secondary data; Asia; large database

It is widely acknowledged that there is great potential for utilizing databases in health care systems in developed countries.¹ Modern health care databases have two characteristics that make them attractive to epidemiologists who work on ‘real-world’ problems: routinely conducted data collection and perpetual electronic data storage. Pharmacoepidemiology and health services research have long taken a leading role in utilizing these kinds of databases in epidemiology, and the emerging research area called comparative effectiveness research (CER) has recently begun to utilize health care databases to weigh the benefits and harms of different interventions.

Health care databases are typically used by researchers as secondary data sources, or existing data generated for a purpose different from the research activity for which it is being used.² There are pros and cons involved in the use of secondary data. From a beneficial aspect, generation of new data is often expensive and time-consuming; further, collection of new data may raise concerns about privacy and unwanted disclosure of data.² Using secondary data for epidemiologic research can therefore save time and resources if the data meet the study purpose. However, while there are great advantages to the use of secondary data, bias in the database, which can be introduced through confounding, missing data, and misclassification, is a great threat to the validity of epidemiologic research.

To address these issues, propensity score matching, sensitivity analysis, and instrumental variable analysis are used to assess and control for bias in observational CER.³ Before applying these novel epidemiologic methods, it is crucial to check the potential for and magnitude of bias by using real-world data. As each database has its own specific purpose (e.g., patient management or health insurance claims), researchers using the databases for secondary purposes sometimes do not know the general characteristics of the databases. The descriptive information provided by databases is therefore important for conducting, analyzing, and interpreting secondary data.

Tanihara analyzed the database of health insurance claims (HIC) of the National Health Insurance Organization of Kumamoto Prefecture in May 2010.⁴ This large database included a total of 3.8 million diagnoses, comprising all computerized health insurance claims in the prefecture for inpatient or outpatient medical care and DPC/PDPS (diagnosis procedure-combination per-diem payment system). Diagnoses in the database were coded in accordance with the International Statistical Classification of Diseases and Related Health Problems, 10th Revision, but some diagnoses were left uncoded. These uncoded diagnoses are often excluded from analyses, which might introduce bias.

Tanihara stated that “the main problem of uncoded diagnoses is that the investigation of large administrative databases using HICs in Japan may have been biased if there was a tendency for one or more specific diagnoses to be uncoded more often than others”. He tackled this issue, finding that the overall proportion of uncoded diagnoses was 9.6% and that proportions differed by HIC type and disease category. Among HIC types, DPC/PDPS included more uncoded diagnoses (14.6%) than outpatient (9.3%) and inpatient (10.9%) records. Among disease categories, congenital malformation, deformations, and chromosomal abnormalities (outpatients; 19.5%); injury, poisoning, and certain other consequences of external causes (inpatients; 19.7%); and neoplasms (DPC/PDPS; 77.9%) showed high percentages of uncoded diagnoses. These findings indicate that Japanese health statistics based on computerized health insurance claims might be biased if we exclude uncoded diagnoses from the analysis.

Rapid economic growth in the Asia-Pacific region has changed population lifestyle and the environment in which we live over the past decade. Non-communicable diseases (NCDs), such as cancer, cardiovascular disease, and diabetes, are now emerging threats to public health in this region, and effective prevention measures for NCDs are needed. The analysis of ‘real-world’ data is recognized as the best means for development of an evidence-based health policy.

Address for correspondence. Prof. Yoshitaka Murakami, PhD, Department of Medical Statistics, Toho University, 5-21-16 Omori-Nishi, Ota-ku, Tokyo 143-8540, Japan (e-mail: yoshitaka.murakami@med.toho-u.ac.jp).

Researchers and public health officers in Japan have drawn attention to the national integrated database of medical receipts and health checkup information, called the 'National Database'. However, access to this database is still restricted under the current government. Miyagawa et al. analyzed part of this database (212 000 residents) and estimated age- and sex-specific prevalence, rates of treatment, and control of cardiovascular risk factors among adults in Shiga prefecture.⁵ Age- and sex-specific prevalence is a fundamental epidemiologic measure that describes health status by region, and stable estimates can only be obtained from large databases. We hope that descriptive epidemiology involving large databases will become more common in the future. We also encourage researchers to utilize large health care databases available in Asian countries.

ACKNOWLEDGMENTS

Conflicts of interest: None declared.

REFERENCES

1. Schneeweiss S. Learning from big health care data. *N Engl J Med*. 2014;370:2161–3.
2. Olsen J. Chapter 23. Using secondary data. In: Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology* third edition. New York: Lippincott Williams & Wilkins; 2008. p. 481–91.
3. Sox HC, Goodman SN. The methods of comparative effectiveness research. *Annu Rev Public Health*. 2012;33:425–45.
4. Tanihara S. The proportion of uncoded diagnosis in computerized health insurance claims in Japan in May 2010 according to ICD-10 Disease Categories. *J Epidemiol*. 2014;24:392–396.
5. Miyagawa N, Murakami Y, Okayama A, Kakuno F, Miura K. Prevalence, treatment, and control of cardiovascular disease risk factors among adults in Shiga Prefecture: Analysis using receipt and health checkup information database in Japan. *Nihon Koshu Eisei Zasshi*. 2014;61:333–41 (in Japanese).