Editorial

Estimating Disease Prevalence and Incidence Using Administrative Data: Some Assembly Required

The prevalence and incidence of a disease are among the most fundamental measures in epidemiology. Prevalence is a measure of the burden of disease in a population in a given location and at a particular time, as represented in a count of the number of people affected. Counts of the number of people with a disease are required to plan appropriately for their healthcare needs. For example, large numbers of patients with rheumatoid arthritis will necessitate training and staffing with more rheumatologists than if the number of patients with the disease was low; similarly, having many patients with osteoarthritis will require more orthopedic surgeons and surgical beds.

Prevalence estimates, which adjust these counts to the size of the source population, are also useful clinically in providing context for diagnostic decision making. Knowing that coronary artery disease is much more common than myocarditis is helpful in evaluating patients with anterior chest pain. Prevalence may also be used to compare disease burden across locations or time periods. However, because prevalence is determined by not only the number of persons affected but also their survival, prevalence is a less useful measure than incidence rates in studies of etiology. Incidence rates represent the number of new cases of disease among the number of susceptible persons in a given location and over a particular span of time. The primary value of incidence rates is in studies of disease etiology, by comparing how the rates vary among different subgroups or with different exposures.

To provide prevalence and incidence rate estimates that are both reliable and generalizable, studies must include a sample large enough to capture most (if not all) cases and sufficiently distributed, both geographically and sociologically, to be representative of the general population. With uncommon diseases, including most autoimmune rheumatic diseases, the challenge is multiplied because cases are fewer and harder to find. These factors necessitate surveys of even larger populations to achieve stable estimates (as well as longer durations of observation for estimates of incidence), which in turn increase the cost, time, and effort involved in executing such studies. Because of these issues, studies using primary data collection to determine the prevalence and incidence of diseases such as systemic lupus erythematosus (SLE) are not common[1]. Understandably, investigators have sought ways to circumvent these issues while still obtaining valid and reliable estimates. In many ways, administrative data that include diagnosis codes fit the bill.

Administrative data are data collected for monitoring, reimbursement, or regulatory purposes, most often by government agencies or insurers, and not primarily for research purposes. However, because administrative data often cover large proportions of the population (with near universal coverage in Canada) and systematically collect similar data elements over years, administrative data are an attractive resource for epidemiological studies. Administrative data have previously been used to obtain estimates of the prevalence or incidence of SLE in Canada, Denmark, Finland, Hong Kong, Iceland, Italy, Norway, Sweden, Taiwan, and the United States[2,3,4,5,6,7,8,9,10,11].

Studies of disease prevalence and incidence involve 3 main activities: assembling the cohort to study; sorting people into affected and unaffected groups (case ascertainment); and counting the number affected. Most methodological work in the use of administrative data to estimate disease prevalence and incidence has focused on the validity of case ascertainment in these data sources[12]. Coding errors, "rule-out" diagnoses, and limits in the number of diagnoses included in a dataset can contribute to errors in administrative data, and the accuracy may vary from data source to data source. Although counting may seem straightforward, capture-recapture methods have been developed to test whether there may be a substantial undercount and to estimate the number of missing cases[13]. In contrast, relatively little attention has been given to how assembly of the cohort may affect the rates.

See Study period affects SLE prevalence/incidence rate, *page 1334*

The type of data source (e.g., hospitalization, outpatient billing), the geographic locations, age or other demographic limits, and number of years of data to include are the main elements to consider in assembling a cohort. In this issue of *The Journal*, Ng and colleagues studied whether estimates of the prevalence and incidence of SLE were sensitive to the number of years of administrative data examined[14]. Estimates of SLE prevalence and incidence were based on both hospital discharges and physician billing codes. The authors defined prevalent cases as those ever coded as having SLE, and patients maintained the diagnosis until death; incident cases were labeled at the first occurrence of a diagnosis of SLE in the database. Using a 15-year period as the reference standard, the authors demonstrated that examining successively shorter time periods resulted in substantially lower prevalence estimates. For example, the prevalence was 46/100,000 when 5 years of data were examined, compared to 60/100,000 when the full 15 years of data were examined. The undercount presumably represents persons who at one time had a healthcare encounter with a diagnosis of SLE but were not recorded as such when a shorter time window was used. Conversely, incidence was higher with shorter time windows compared to the full 15-year period, because in the shorter window prevalent cases were misclassified as new diagnoses. The incidence was estimated to be 8/100,000 with 5 years of data, but only 5.6/100,000 with 15 years of data. The authors suggest that more than 5 years of data are likely needed to avoid these issues and provide valid estimates.

This study highlights the importance of cohort assembly in studies using administrative data, and in particular, the need to think carefully about not only the data included in the study, but also about the data omitted. This study shows that the omitted data can have a major influence when studies examine only a small segment of time. Administrative data are susceptible to these effects because all persons with the disease are not identified continuously, but rather flagged only when they use healthcare services, and because incident cases cannot be distinguished from prevalent cases in a given year, but only by looking in previous years.

Despite the importance of these observations, some caveats remain. Ng and colleagues assessed period prevalence over 15 years, which is quite long. Point prevalence studies or studies with period prevalences of 5 years or less are more common, because those studies provide more contemporary estimates. A potential limitation of using 10-year or 15-year periods is that they may not accurately reflect the current status of the disease, particularly if temporal changes have occurred because of the introduction of a more sensitive diagnostic test or a remission-inducing treatment, for example. The underestimate of prevalence during successively shorter time windows noted by Ng and colleagues was due to previously diagnosed patients not having a healthcare encounter coded as SLE in those windows. The patients' absence from the shorter time windows indicates they are not currently consuming healthcare resources related to SLE. If a major objective of prevalence estimation is to aid healthcare planning, these patients may be less relevant. In addition, studies of incidence rates based on administrative data tend to be of limited value in understanding disease etiology, because these sources generally lack data on potential exposures. Comparisons are often limited to demographic characteristics and geography, which can provide only crude suggestions about disease etiology. One exception to this generalization is that temporal trends in incidence rates may provide clues to etiology if these can be linked to trends in other data sources.

Despite the greater availability of administrative data, researchers may not have access to a decade or more of data. Can the principles outlined by Ng and colleagues be incorporated in studies that include only a few years of data? In prevalence studies, a short window will capture only a proportion of the all patients, for example those who happen to be hospitalized or who have a physician visit coded as SLE-related in the years included. If the proportion of patients who are hospitalized or who are treated in a given year can be obtained from other sources or from clinic data, an estimate of the number of prevalent patients in the population can be derived by dividing the number observed in the administrative dataset by the proportion hospitalized (or treated) per year. This inflation adjustment results in prevalence estimates that are quite accurate, particularly when the estimate of the proportion hospitalized per year is also population-based[15]. For incidence studies, a 1- or 2-year lag period can be included, so that any patients who appear in this period are not counted as incident. Incident cases would be counted only among those who were known to have at least some years of followup without a prior qualifying diagnosis. This lag helps to approximate observational studies of incidence by assembling a cohort "without disease" at entry. Consideration of these issues will result in studies with more valid estimates of disease prevalence and incidence.

MICHAEL M. WARD, MD, MPH,
Intramural Research Program,
National Institute of Arthritis and
Musculoskeletal and Skin Diseases,
National Institutes of Health,
Bethesda, Maryland, USA

Supported by the Intramural Research Program, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health.

Address correspondence to Dr. M. Ward, Building 10 CRC,
Room 4-1339, 10 Center Drive, Bethesda, MD 20892-1468.
E-mail: wardm1@mail.nih.gov

REFERENCES

1. Lim SS, Drenkard C, McCune WJ, Helmick CG, Gordon C, Deguire P, et al. Population-based lupus registries: advancing our epidemiologic understanding. Arthritis Rheum 2009;61:1462-6.
2. Bernatsky S, Lix L, Hanly JG, Hudson M, Badley E, Peschken C, et al. Surveillance of systemic autoimmune rheumatic diseases using administrative data. Rheumatol Int 2011;31:549-54.
3. Voss A, Green A, Junker P. Systemic lupus erythematosus in Denmark: clinical and epidemiological characterization of a county-based cohort. Scan J Rheumatol 1998;27:98-105.
4. Helve T. Prevalence and mortality rates of systemic lupus erythematosus and causes of death in SLE patients in Finland. Scand J Rheumatol 1985;14:43-6.
5. Mok CC. Epidemiology and survival of systemic lupus erythematosus in Hong Kong Chinese. Lupus 2011;20:767-71.
6. Gudmundsson S, Steinsson K. Systemic lupus erythematosus in Iceland 1975 through 1984. A nationwide epidemiological study in an unselected population. J Rheumatol 1990;17:1162-7.
7. Govoni M, Castellino G, Bosi S, Napoli N, Trotta F. Incidence and prevalence of systemic lupus erythematosus in a district of north Italy. Lupus 2006;15:110-3.
8. Nossent HC. Systemic lupus erythematosus in the Arctic region of Norway. J Rheumatol 2001;28:539-46.
9. Stähl-Hallengren C, Jönsen A, Nived O, Sturfelt G. Incidence studies of systemic lupus erythematosus in southern Sweden: increasing age, decreasing frequency of renal manifestations and good prognosis. J Rheumatol 2000;27:685-91.
10. Chiu YM, Lai CH. Nationwide population-based epidemiological study of systemic lupus erythematosus in Taiwan. Lupus 2010;29:1250-5.
11. Furst DE, Clarke AE, Fernandes AW, Bancroft T, Greth W, Iorga SR. Incidence and prevalence of adult systemic lupus erythematosus in a large U.S. managed-care population. Lupus 2013;22:99-105.
12. Widdifield J, Labrecque J, Lix L, Paterson JM, Bernatsky S, Tu K, et al. A systematic review and critical appraisal of validation studies to identify rheumatic diseases in health administrative databases. Arthritis Care Res 2013 Feb 22 [E-pub ahead of print].
13. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995;17:243-64.
14. Ng R, Bernatsky S, Rahme E. Observation period effects on systemic lupus erythematosus incidence and prevalence estimation from Quebec administrative data. J Rheumatol 2013;40:1334-6.
15. Ward MM. Estimating rare disease prevalence from administrative hospitalization databases. Epidemiology 2005;16:270-1.

J Rheumatol 2013;40:1241–3; doi:10.3899/jrheum.130675