Special Communication

Institutional and Surveillance Database use in Epidemiologic Research in Developing Countries: Revisiting Some Limitations

Muhammad Younus[1], Azfar-e-Alam Siddiqi[2], Bilquis Sana Khan[3], Amy L. Steffey[4]

Department of Epidemiology[1,2], College of Human Medicine, Michigan State University, East Lansing, Michigan, USA, Department of Community Health Sciences[3], The Aga Khan University, Karachi, Pakistan, School of Public Health[4], University of Michigan Ann Arbor, Michigan, USA.

Background

Disease surveillance in public health is centuries old-the earliest surveillance systems used to monitor population health trends which were vital statistics records developed in Europe in the 18th Century.[1] Surveillance for diseases and health-related events has been an essential component of public health systems in developed countries for decades.[2] These surveillance systems not only monitor and evaluate health-related events, but also support epidemiological research. Data based on surveillance or disease registries are used to estimate population disease burdens and may also be used to study associations between exposures and outcomes (hypothesis testing). In countries with highly developed comprehensive surveillance systems, disease reporting to healthcare authorities range from voluntary to mandatory reporting by law.[3] Despite the advancement in surveillance system attributes, potential for various sources of biases in reporting cannot be completely prevented. Surveillance bias refers to any systematic, exposure of interest or outcome-related activity that exists for some individuals or groups and not for others.[4] Sources of such bias include differences in disease reporting practices and inconsistencies in case definitions and diagnoses among healthcare professionals and medical institutions, along with differences in the health seeking behaviour of populations.

In Pakistan and other developing countries where centralized, population-based comprehensive disease surveillance systems are not well developed, 2 researchers and healthcare providers have attempted to run small-scale, focused, disease-specific surveillance projects, some with good success. But the failure to follow one standard approach and the resulting variation, make this surveillance data even more vulnerable to bias.

Limited resources often restrict primary data collection for estimating magnitude of disease in populations. As a result, epidemiologists and clinicians, particularly in developing countries, are often forced to utilize institutional databases such as hospital records for estimating population disease burdens and testing research hypotheses. Like surveillance data, the institutional databases have their own limitations.

In this report, we will highlight the types of biases related to disease surveillance and institutional databases that when present, compromise the integrity of the data and affect the validity of research findings conducted using these data. It should be noted that these types of biases are not mutually exclusive, but somewhat interrelated.

Selection bias

A patient's visit to an outpatient or emergency department of a healthcare facility is a prerequisite for disease reporting to occur. A significant proportion of individuals, usually with subclinical to mild disease symptoms, are not diagnosed or do not even seek treatment. In countries like Pakistan, this is further compounded by over the counter distribution of prescription medicines, which promotes self-medication and provides the patient an alternative to seeking a qualified healthcare professional. For example, if a research project utilizing institution-based surveillance data or hospital records aims to estimate the disease burden related to foodborne illnesses such as salmonellosis or campylobacteriosis in a population, the results will underestimate the actual magnitude of the disease[5], and may also be restricted to the more serious cases.

Additionally, health seeking behaviour varies among populations and is partly determined by the educational level and health knowledge of communities, which could significantly influence disease reporting rates among populations.

Referral bias

Physicians and healthcare facilities may attract patients with certain diseases, health-related conditions, or exposures from beyond their catchment areas, resulting in a higher proportion of patients with those diseases or conditions in a hospital database. Applying estimates generated from such data to the target population could result in the overestimation of a disease in the population.[6] Similarly, if data from a laboratory or a tertiary care center is used to obtain the rate of positive results for certain infections, the results may be biased. In the developing world, especially in peri-urban areas, patients with positive test results are referred to major diagnostic centers for confirmatory diagnosis. For example, primary care physicians in remote locations typically refer patients with suspected multidrug-resistant tuberculosis (MDR-TB) to physicians in tertiary care hospitals for confirmation and further management. As a result, laboratory databases for these hospitals would show an increase in the positive predictive value of the test performed and also a higher proportion of MDR-TB among all TB patients. Estimates based on these hospitals records would overestimate the magnitude of the disease in the population.

Berkson's bias

The validity of epidemiologic research findings utilizing hospital data may be uncertain due to Berkson's bias. In 1964, Joseph Berkson explained that the relative frequency of disease in hospitalized individuals is inherently biased when compared to the population served by the hospital. This phenomenon is attributed to the way probabilities of hospitalization combine in patients with severe medical conditions or with more than one disease.[6] Berkson's argument applies in particular to hospital-based case-control studies where the associations between risk factors and diseases are studied. If, for example, obese people who are also hypertensive have a higher probability of being hospitalized than people with hypertension alone, a spurious association between obesity and anti-hypertensive drugs can be found.

Misclassification of exposure or outcome

The International Classification of Diseases (ICD) is the standard diagnostic classification used to categorize diseases and other health-related events such as those contained in death certificates and hospital records. The ICD aids in the storage and retrieval of diagnostic information for clinical and epidemiological research purposes and allows for compilation and comparison of national morbidity and mortality statistics.[7] The ICD is widely used as a disease classification system in the developed world and has helped standardize diagnoses. However, this data system is not foolproof; it does not eliminate misclassification[8], which is compounded by differences in coding accuracy across institutions.[9] Use of ICD data for research purposes in developing countries is hindered by variability of its use, in addition to the limitations mentioned above. Lack of widespread and consistent use of ICD in data abstraction may lead to misclassification of exposures or outcomes due to inconsistent case definitions. Therefore, limitations should be noted when making comparisons between healthcare settings or generalizing study results based on databases that do not use ICD codes. Adherence to ICD coding when abstracting data does however have the potential of reducing the variability in diagnoses and can improve the specificity of reporting and comparability between data sources.

Conclusions

In recent years, with the growing biomedical research field in developing countries including Pakistan, epidemiologists and physicians are increasingly using surveillance and institutional databases for estimating disease burdens and comparing rates among different populations. We do not intend to criticize epidemiologic methods or databases in this report; our purpose is to highlight points that will facilitate a realistic appraisal of surveillance database use in epidemiologic research that may enhance the methodological aspects of future research.

There are no easy solutions to eliminating inherent biases related to surveillance or hospital-based data. The objective is to recognize the effects of reporting biases and minimize misinterpretation of results by readers. We recommend researchers discuss the limitations of databases and acknowledge possible sources of bias when reporting results based on surveillance data or medical record audits.

References

1. Ritz B, Tager I, Balmes J. Can lessons from public health disease surveillance be applied to environmental public health tracking? Environ Health Perspect 2005; 113:243-9.
2. Akhtar S, White F. Animal disease surveillance: prospects for development in Pakistan. Rev Sci Tech 2003; 22:977-87.
3. Roush S, Birkhead G, Koo D, Cobb A, Fleming D. Mandatory reporting of diseases and conditions by health care professionals and laboratories. JAMA 1999; 282:164-70.
4. Chaffin M, Bard D. Impact of intervention surveillance bias on analyses of child welfare report outcomes. Child Maltreat 2006; 11:301-12.
5. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee J, Shapiro C, et al. Food-related illness and death in the United States. Emerg Infect Dis 1999; 5:607-25.
6. Schwartzbaum J, Ahlbom A, Feychting M. Berkson's bias reviewed. Eur J Epidemiol 2003; 18:1109-12.
7. World Health Organization (WHO). International Classification of Diseases (ICD). [online] [cited 2007 January 15]. Available from: RUL: http://www.who.int/classifications/icd/en/.
8. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res 2005; 40:1620-39.
9. Golomb MR, Garg BP, Saha C, Williams LS. Accuracy and yield of ICD-9 codes for identifying children with ischemic stroke. Neurology 2006; 67:2053-5.