INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 78 (2009) 417–424

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Differences between notifiable and administrative health information in the spatial–temporal surveillance of enteric infections

N. Yiannakoulias[a,*], L.W. Svenson[b]

[a] School of Geography and Earth Sciences, General Science Building Room 204, McMaster University, 1280 Main Street West, Hamilton, ON, Canada L8S 4K1

[b] Surveillance and Environmental Health, Alberta Health and Wellness, Canada

ARTICLE INFO

Article history:
Received 25 July 2007
Received in revised form
29 December 2008
Accepted 30 December 2008

Keywords:
Surveillance
Communicable diseases
Data sources

ABSTRACT

Purpose: The purpose of this study is to compare the spatial and temporal information generated from two distinct health data sources available for the surveillance of intestinal infections associated with *Escherichia coli* O157:H7.

Methods: Our study area is the province of Alberta, Canada. Data are from two sources: a fee-for-service administrative health data system and a notifiable disease data reporting system. The study period is between 1999 and 2005. We compare the systems by observing correlations in the infections over time, the variability in the overall distribution of cases (as measured by a geographic dissimilarity index), and the relative locations of spatial–temporal clusters of infection.

Results: Our results indicate considerable variability in information generated from these two systems. The geographic distribution of cases varies considerably, with annual indices of dissimilarity suggesting considerable variation in the geographic distribution of cases throughout the study period ($D = 0.445$). The temporal patterns identified by these two sources of information are negatively correlated (-0.40, $p < 0.001$). Notifiable disease clusters occur in the summer in southern regions of the province, whereas cases identified from administrative health data system cluster in the winter season, and further to the north.

Conclusions: Notifiable disease data may suffer from selection bias; administrative health data may be insufficiently precise without laboratory confirmation. Our results illustrate differences in the spatial and temporal information generated from these two systems of case identification. Future surveillance of gastrointestinal illness of infectious origin may benefit from case ascertainment algorithms based on both sources of data.

1. Introduction

Considerable research has gone into the design, application and evaluation of administrative health data systems in recent years. New data standardization models [1] and case identification strategies based on multiple data sources (such as medical claims, inpatient visits and emergency room encounters) [2–3] have enhanced the application of these data systems in recent years. On the other hand, data quality issues [4], privacy concerns [5] and data

* *Corresponding author*. Tel.: +1 905 525 9140x20117.
E-mail address: yiannan@mcmaster.ca (N. Yiannakoulias).

1386-5056/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2008.12.006

linkage strategies [6–7] remain important considerations associated with their use. Nonetheless, routinely collected administrative data remain an important source of inexpensive and often comprehensive disease information [8–9] suited to retrospective study [10] and disease surveillance [11].

Much research has critically examined the value of administrative health data for identifying cases of chronic and acute conditions [12–15]. For some time there had been less interest in using administrative health data systems for research in and surveillance of communicable diseases. This is likely because such sources are deemed to lack the necessary specificity, especially for communicable diseases which do not present symptoms that can be easily diagnosed without laboratory verification. However, the need for outbreak detection and investigation has brought about a new perspective on the data resources currently available for routine monitoring population health, and the adoption of a new 'syndromic' surveillance paradigm. Syndromic surveillance refers to the identification of population patterns and indicators of disease based on cases defined through non-confirmed (e.g., laboratory) sources [16]. The syndromic perspective emphasizes the importance of timeliness and sensitivity, and is well suited to the detection of changes in disease that are of public health importance—such as newly emerging pathogens and bioterrorism.

As population-based administrative data gain acceptance as a source of health information, serious attention is now being paid to their potential as a source of information that supplements laboratory confirmed notifiable disease data [16–17]. Although notifiable disease data can be used to identify the presence of disease-causing pathogens, these sources are often too sparse to detect outbreaks at high spatial and temporal resolutions (such as at the neighbourhood level, or at daily time frames) and may suffer from various forms of selection bias [18,19]. On the other hand, while administrative health data do not typically contain information on the presence or type of infectious agent, they are often comprehensive, and conducive to high-resolution spatial and temporal surveillance.

Spatial–temporal methodologies are critical to routine disease surveillance, as they provide information about the 'where' and 'when' of disease. In many jurisdictions, both notifiable data and administrative health data are available for spatial–temporal surveillance of gastrointestinal infections. Although some research has compared the validity of these sources, and other research has evaluated their capacity to detect known disease outbreaks in a timely manner [20], little work has compared the spatial–temporal information that these different sources of data offer, especially at large geographic scales.

This study compares information derived from laboratory confirmed *Escherichia coli* O157:H7 infections contained in a notifiable disease registry with information derived from a population based fee-for-service medical claims system. Our analysis compares the temporal, spatial and spatial–temporal information generated by these two data systems. By comparing this information, we hope to contribute to a greater understanding of their relative contributions to surveillance activities, and help inform the planning of public health surveillance activities in jurisdictions where similar data systems exist.

2. Methods

2.1. Data

We use data from between 1999 and 2005 from the province of Alberta, Canada. Our study compares two different sources of disease surveillance information: data from a notifiable disease registry system and data from a fee-for-service medical claims system. The Communicable Disease Reporting System (CDRS) maintained by the provincial health ministry contains information on a variety of notifiable diseases reported in Alberta. The CDRS database is populated with information obtained through patient specimen samples taken at the discretion of medical practitioners (e.g., physicians). Testing laboratories submit results to the health ministry, and upon receipt of these results, records management personnel submit a Notifiable Disease Report (NDR) in support of the laboratory findings. A case is confirmed when an NDR report and positive laboratory notification are submitted to the provincial health ministry. For notifiable diseases, submission of these data is mandated by law.

Over 99% of the provincial population is covered by the provincial public insurance system which covers the majority of medically necessary services. This includes a wide range of services including primary care, secondary care, acute care and other services. Like other Canadian provinces, the provincial government acts as an insurer, and most physicians are remunerated by the provincial government on a fee-for-service (FFS) basis. The FFS payment system necessitates the creation and maintenance of an electronic database that contains information on the services provided, the conditions associated with each service (usually recorded as an ICD-9 or ICD-10 code) the provider of the service (including the medical specialty of the provider) and the site at which the service was performed. Since submission of records to the provincial health ministry is a pre-condition of payment for a service, compliance is very high, and 85% of the service records are received within 2 weeks. The records in the database can be linked to other data systems, including most records in the CDRS system, using a unique personal health number associated with registration in the public health system. Both the FFS and CDRS data can also be linked to a population insurance registry, which contains information on place of residence.

2.2. Case selection criteria

Positive laboratory confirmation of *Escherichia coli* O157:H7 (*E. coli*) infections are used to identify cases in the CDRS system. FFS cases are identified based on a combination of diagnostic codes, provider information and dates of service. To ensure that our results are not an artefact of a particular case selection algorithm, we select cases based on several combinations of criteria. First, three sets of ICD-9 diagnostic codes associated with *E. coli* infections were considered: 008.x, 009.x and 787.9, with various permutations of the 4th digit considered. For the 009.x and 787.9 diagnostic codes, two or more visits with these