Journal of Hospital Infection 89 (2015) 267–270

Available online at www.sciencedirect.com

Journal of Hospital Infection

journal homepage: www.elsevierhealth.com/journals/jhin

Role of data warehousing in healthcare epidemiology

D. Wyllie[a],*, J. Davies[b]

[a] Public Health England Academic Collaborating Centre, John Radcliffe Hospital, Oxford, UK
[b] Oxford NIHR BRC Informatics Programme, Department of Computer Science, University of Oxford, Oxford, UK

ARTICLE INFO

Article history:
Received 25 November 2014
Accepted 6 January 2015
Available online 29 January 2015

Keywords:
Data management
Epidemiology
Healthcare data

SUMMARY

Electronic storage of healthcare data, including individual-level risk factors for both infectious and other diseases, is increasing. These data can be integrated at hospital, regional and national levels. Data sources that contain risk factor and outcome information for a wide range of conditions offer the potential for efficient epidemiological analysis of multiple diseases. Opportunities may also arise for monitoring healthcare processes. Integrating diverse data sources presents epidemiological, practical, and ethical challenges. For example, diagnostic criteria, outcome definitions, and ascertainment methods may differ across the data sources. Data volumes may be very large, requiring sophisticated computing technology. Given the large populations involved, perhaps the most challenging aspect is how informed consent can be obtained for the development of integrated databases, particularly when it is not easy to demonstrate their potential. In this article, we discuss some of the ups and downs of recent projects as well as the potential of data warehousing for antimicrobial resistance monitoring.

© 2015 Published by Elsevier Ltd on behalf of the Healthcare Infection Society.

Introduction

Healthcare epidemiology, a branch of epidemiology concerned with the detection, control, and prevention of adverse events in the health economy, has gained prominence in recent years.[1] This is attributable partly to a desire to learn more about the determinants of morbidity, mortality and cost in modern healthcare, and partly to an expectation that continuous quality monitoring and benchmarking can be built into efficient management systems.[2]

Data warehousing is a process by which information can be shared efficiently; the level on which it is shared may be an organization, a region, a network, or a nation.[3] Information in the data warehouse may be in the form of a single definitive record, as occurs with integrated electronic patient care systems in some hospitals. This 'top down' strategy has many attractions, but can be hard to implement and may be impractical when different systems play key roles not available in a core system.[3] Another common scenario uses multiple independent systems, resolving the problem posed by systems that cannot readily exchange information with each other by making the different systems contribute instead to a common 'information pool', thus allowing access to data across the organization as a whole. Inconsistencies may exist between data received from the different systems, but these can be resolved in order to generate a consistent 'single source of data' to inform policy making.[3] This latter scenario is widespread in hospitals, where dozens or hundreds of independent systems may be in use.[4,5] Some of these systems may contain very large amounts of information, such as databases containing laboratory tests or those tracking patient movements, others being restricted to much smaller patient populations. The quality of these small data sets, sometimes called 'long tail data', may be very high, but they may be ignored in data warehouses owing to the cost of integrating them.[6]

* Corresponding author. Address: Public Health England Academic Collaborating Centre, John Radcliffe Hospital, Headley Way, Oxford OX3 9DU, UK. Tel.: +44 (0)1865 220860.

E-mail address: david.wyllie@ndm.ox.ac.uk (D. Wyllie).

http://dx.doi.org/10.1016/j.jhin.2015.01.005
0195-6701/© 2015 Published by Elsevier Ltd on behalf of the Healthcare Infection Society.

As in other fields, the information generated in healthcare is increasing rapidly. Widespread use of electronic patient records, routine recording of free-text communications (such as discharge summaries), real-time patient tracking, continuous monitoring of patient physiology, cross-sectional imaging, telehealth, and human and microbial genomics are all contributing. Integrating these data to create what is called 'big data' presents technical and strategic challenges; if successful, the integrated data will relate both to outcomes of interest (such as mortality, complications, or microbial spread) and to a wide range of risk factors.

'Big data' is a concept defined as electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware; nor can they be easily managed with traditional or usual data management tools and methods.[7] The range of problems with big data has been expressed as 'the four Vs': velocity, variety, volume, and validity.[7] These problems apply to a large extent to all aspects of data warehousing. In this article, we will illustrate problems and solutions associated with each of them, using examples from healthcare epidemiology. We will also discuss the ethical and social issues associated with large-scale information integration, as well as the feasibility of using big data for day-to-day monitoring of antimicrobial resistance.

Successes and challenges in data warehousing for healthcare epidemiology

Velocity

The first challenge concerns the rate at which data are accrued and the interval between accrual and analysis. Although many decisions (such as about antibiotic policy) are made on the basis of historical data sets that may have been gathered some time before analysis, other problems require much more rapid information synthesis and reporting. This may require particular kinds of storage arrangements. An example of 'high speed' data is use of emergency room monitoring to detect clinical deterioration, by integration of data from multiple sources to produce a single measure of need for increased care.[8] Velocity is also important in syndromic surveillance systems designed to detect clusters indicative of point-source outbreaks or deliberate pathogen release.[9]

Variety

A second challenge concerns the diversity of data sources used to assess the probability of an event of interest. One example is found in the use of research articles, satellite data on climate, historical and contemporary laboratory reports, and crowd-sourced reports of diagnoses to produce maps of disease.[10] These data are highly disparate, but, by combining them, useful inferences may be drawn. Another example concerns post-marketing surveillance of pharmaceuticals, where multiple data sources are integrated to inform assessments of drug safety.[11]

Volume

The third issue concerns data volume. A large UK hospital with coverage of about 0.7% of England has about 2 TB (terabytes) of data stored in relational databases, excluding radiology and genetic data, whereas the US Kaiser Permanente health system stores about 5000 times more data, in excess of 10 PB (petabytes).[5,7] Large databases are also being assembled in Europe, for example by the English National Health Service's Health and Social Care Information Centre (http://www.hscic.gov.uk), and for microbial surveillance initiatives by Public Health England. Even the very large databases accrued in healthcare are small compared with data volumes generated in physical and astronomical sciences.[12] The increasing volume of data stored obviously has an impact on the hardware and software required, leading to the emergence of new technical solutions.[7] Nevertheless, there is a substantial cost to this, which can run into tens or hundreds of thousands of euro (€) per annum. The key advantages of large data volumes include wide area coverage, and the ability to detect small effects on rare outcomes. A study by Freemantle *et al*. exemplifies this. They investigated mortality at weekends in admissions to all English hospitals, and in a group of 254 US hospitals.[13] Admission at the weekend was associated with a significant increase in mortality that could not be explained by altered case-mix in the data sets available. These results have prompted a review of care delivery models in England. Another example is provided by the work of Shorr *et al*.[14] Analysing 62 US hospitals over a four-year period, they identified 5975 patients with a clinical diagnosis of pneumonia that was supported by laboratory evidence. Of this cohort, 837 (14%) had pneumonia due to meticillin-resistant *Staphylococcus aureus* (MRSA). Even though this represents fewer than four cases per hospital per annum, the authors developed and validated a risk score for MRSA pneumonia, which they suggested might be used to restrict the use of anti-MRSA antimicrobial agents in high-risk areas. It seems unlikely that such a study could have been performed without use of a wide-area database. A third example using very large data sets is the detection of rare adverse events following drug or vaccine licensing. The rationale is that the impact of the drug or vaccine after licensing may differ from that before licensing, or that rare side-effects may remain undetected before licensing. This requires analysis of diverse data sources.[11]

Validity

Are the data in the warehouse accurate enough for the intended use? Factors compromising validity would include the various kinds of biases that are known to compromise epidemiological studies, such as selection bias and misclassification bias, due to problems with the information used to identify the patients, the outcomes and the covariates of interest.[15] A contemporary example illustrating validity problems is provided by Google Flu Trends, an algorithm which used search terms entered into Google to predict influenza incidence.[16] Search terms associated with influenza incidence, determined by the US Centers for Disease Control and Prevention laboratory-based surveillance system, were identified and a model built that predicted influenza incidence based on these data. After an initial period of success, the model substantially overestimated influenza incidence. Two factors may have been responsible for this decline in performance: a period of public concern about influenza; and an alteration in the way that Google 'suggests' search terms to its users.[16] This emphasizes the importance, for valid epidemiological

Download English Version:

https://daneshyari.com/en/article/6122099

Download Persian Version:

https://daneshyari.com/article/6122099

Daneshyari.com