Can J Diabetes 39 (2015) 235–238

Contents lists available at ScienceDirect

Canadian Journal of Diabetes

journal homepage: www.canadianjournalofdiabetes.com

Perspectives in Practice

Clinical Diabetes Research Using Data Mining: A Canadian Perspective

Baiju R. Shah MD, PhD[a,b,c,*], Lorraine L. Lipscombe MD, MSc[a,b,d]

[a] *Department of Medicine, University of Toronto, Toronto, Canada*
[b] *Institute for Clinical Evaluative Sciences, Toronto, Toronto, Canada*
[c] *Department of Medicine, Sunnybrook Health Sciences Centre, Toronto, Canada*
[d] *Department of Medicine, Women's College Hospital, Toronto, Canada*

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form
11 February 2015
Accepted 23 February 2015

Keywords:
administrative data
data mining
research methods

Mots clés :
données administratives
exploration de données
méthodes de recherche

ABSTRACT

With the advent of the digitization of large amounts of information and the computer power capable of analyzing this volume of information, data mining is increasingly being applied to medical research. Datasets created for administration of the healthcare system provide a wealth of information from different healthcare sectors, and Canadian provinces' single-payer universal healthcare systems mean that data are more comprehensive and complete in this country than in many other jurisdictions. The increasing ability to also link clinical information, such as electronic medical records, laboratory test results and disease registries, has broadened the types of data available for analysis. Data-mining methods have been used in many different areas of diabetes clinical research, including classic epidemiology, effectiveness research, population health and health services research. Although methodologic challenges and privacy concerns remain important barriers to using these techniques, data mining remains a powerful tool for clinical research.

RÉSUMÉ

Avec l'avènement de la numérisation de grandes quantités d'informations et la puissance informatique capable d'analyser ce volume d'informations, l'exploration de données est de plus en plus appliquée à la recherche médicale. Les jeux de données créés pour l'administration du système de santé fournissent une mine d'informations provenant de différents secteurs de la santé, et les systèmes de santé universels à payeur unique des provinces canadiennes signifie que les données sont plus générales et complètes dans ce pays que dans beaucoup d'autres juridictions. L'augmentation de la capacité de lier l'information d'origine clinique, tels les dossiers médicaux électroniques, les résultats des tests de laboratoire et les registres des maladies, a élargi les types de données disponibles pour analyse. Les méthodes d'extraction de données ont été utilisées dans de nombreux domaines de la recherche clinique du diabète, y compris l'épidémiologie classique, la recherche d'efficacité, la recherche sur la santé des populations et des services de santé. Bien que les défis méthodologiques et de confidentialité restent d'importants obstacles à l'utilisation de ces techniques, l'exploration de données reste un outil puissant pour la recherche clinique.

Introduction

Data mining is the process of exploring and analyzing large datasets to find previously unknown relationships or correlations and to summarize the data in novel ways that are both understandable and useful (1). Applications of data mining are common in many industries, such as credit card fraud detection and targeting marketing campaigns. The mining of health data is being applied increasingly in medical research to track healthcare patterns and explore disease hypotheses. Here, we review the unique Canadian role in clinical research using data mining, some applications of data mining in clinical diabetes research and some ongoing challenges facing the application of data mining in health research.

Data mining has become possible because of the advent of big data—the creation and collection of vast amounts of information in a digital format. Digitalized data have dramatically increased in

[*] Address for correspondence: Baiju R. Shah, MD, PhD, G106, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada.
E-mail address: baiju.shah@ices.on.ca

1499-2671/$ – see front matter © 2015 Canadian Diabetes Association
http://dx.doi.org/10.1016/j.jcjd.2015.02.005

volume, velocity and variety. According to IBM, 2.5 exabytes (that's 2,500,000,000,000 megabytes) of data were generated each day in 2012. The types of information being collected have also expanded to include both structured and unstructured data, such as images, videos, sensor outputs and gene sequencings. Major advances in computing power have, therefore, been necessary to allow meaningful and efficient access to and analysis of such large volumes of data. In computer sciences, Moore's law observes that microprocessor complexity doubles approximately every 2 years (2); hence, the average smartphone today has computing power that is several orders of magnitude greater than that of the original IBM personal computer. These 2 parallel technologic advances of the past few decades have been catalysts that have permitted the development of modern data mining.

As in other industries, the digitization of data has been adopted increasingly in the healthcare setting. Administrative and financial information in large institutions such as hospitals has long been managed electronically; but now, clinical information such as radiology images and primary care physicians' records are also becoming digitalized. In all cases, these large datasets are collected for reasons other than analysis, such as for clinical care or administration of the healthcare system. However, they can be secondarily exploited for clinical research purposes. (Data mining techniques are also used to exploit large datasets in other areas of health research, such as genome-wide association studies or functional neuroimaging, but these are beyond the scope of this review.) In the past, observational health research relied on small registries, health surveys or selected cohorts. The use of large secondary databases increases the potential sample size and power for such studies and provides unique opportunities to explore novel questions in unselected population-based cohorts. Furthermore, because these databases are being created anyway, for other purposes, both the costs of doing research and the time taken to obtain results can be substantially lower when using these data than by setting out to collect data prospectively.

Canada's Place in Healthcare Data Mining

Canada's healthcare system is uniquely positioned to develop large databases suitable for population-based research (3). Within each province and territory, virtually every permanent resident is eligible for healthcare services provided by a single payer; therefore, the data collected to administer the system include virtually the entire population. There are very few inpatient and physician services that are outside of the public system, so virtually all healthcare interactions in these sectors can be captured. In addition, each province and territory assigns a unique personal identifier (the health card number) to all residents, which is captured as part of clinical care; therefore, individuals' records from various databases and various healthcare sectors can be linked. Administrative data sources detailing many different healthcare sectors are available (though availability varies among provinces), including health insurance registration data, inpatient and outpatient hospital care, day surgery, emergency department care, physician services, homecare, long-term care, vital statistics, prescriptions and healthcare provider data. These are increasingly being joined by large databases of clinical information, including laboratory tests, healthcare providers' electronic medical records, and cancer- and other disease-based registries. Several Canadian provinces have developed data warehouses, such as Population Data BC, the Manitoba Centre for Health Policy, Ontario's Institute for Clinical Evaluative Sciences and Health Data Nova Scotia. The purpose of these warehouses is to provide a more convenient repository for these databases and advanced expertise in the linkage and analyses of information for research and healthcare evaluation.

Uses of Data Mining for Clinical Diabetes Research

Modern data mining can be used in the same way that population-level data have always been used: for classic epidemiologic studies searching for patterns of and causes of disease in populations. When John Snow became the father of modern epidemiology while trying to understand the 1854 cholera outbreak in London, he was using a 19th century version of big data: a map of Soho plotting the homes of cholera victims. With 21st century digitized healthcare administrative data, we can continue to perform such classic epidemiologic studies, such as comparing diabetes prevalence across provinces (4) or among neighbourhoods within a city (5). These epidemiologic techniques can also be used to detect associations between diabetes and other medical conditions that are not traditional complications of the disease, such as hip fracture, breast cancer or sepsis (6–8). By virtue of being "big," the data available from healthcare administrative sources may be useful for finding associations that would be too difficult to find in smaller datasets. For example, the relationships between gestational diabetes and cardiovascular risk factors and between gestational diabetes and proxies for cardiovascular disease were established with prospective cohort studies and cross-sectional analyses of clinical data. However, only mining of large databases could find an association between gestational diabetes and cardiovascular events because they are so rare in reproductive-age women (9).

Phase IV postmarketing surveillance for adverse events caused by new drugs requires large numbers of patients with a wide variety of medical conditions to be exposed to drugs over the long term. The adverse events are often very rare and not necessarily known or suspected a priori. Hence, data mining is a particularly valuable tool for drug safety and pharmacoepidemiologic research. Data mining was used in many of the studies that were evaluating whether there is a cardiovascular risk associated with thiazolidinediones (10–12) and in studies linking statins and fluoroquinolones with glucose abnormalities (13–15). Data mining also contributes to understanding the safety of older pharmaceutical agents, like sulfonylureas and metformin (16–19). Distributed data networks are being created to enable evaluation of rarer adverse drug effects across very large populations. For example, the Canadian Network for Observational Drug Effects Studies (CNODES) is a pan-Canadian collaboration of researchers that provides information regarding drug safety by combining data from 9 population-based databases using standardized methods (13,20,21).

Another avenue of clinical research in which analysis of secondary data sources is invaluable is effectiveness research—studying whether an intervention works in routine clinical care (which contrasts with efficacy research, which studies whether it works on carefully selected patients in idealized settings such as a randomized trial). Because effectiveness, by definition, requires examination of the real-world care of "regular" patients, routinely collected administrative data are an ideal method for conducting this research. For example, although the use of spironolactone in patients with heart failure reduced mortality in a randomized trial, its use in routine clinical care was associated with an abrupt increase in hyperkalemia-associated mortality (22). Similarly, these data are ideally suited for population health questions because they include information on the entire unselected population. Thus, Canadian researchers have used data mining to explore the relationships among neighbourhood walkability, poverty and diabetes risk (23) and the influence of sex and ethnicity on macrovascular complications of diabetes (24,25).

Data mining can also be used for health services research and evaluation of healthcare system performance. Because these data are often collected for healthcare administrative purposes, they are particularly suited for this type of research. For example, the impact