Review

JOURNAL OF HEPATOLOGY

Secondary analysis of large databases for hepatology research

Philip N. Okafor[1,*], Maria Chiejina[2], Nicolo de Pretis[3], Jayant A. Talwalkar[1,4]

[1]*Division of Gastroenterology and Hepatology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, United States;* [2]*Department of Internal Medicine, Good Shepard Medical Center, Longview, TX 75601, United States;* [3]*Division of Gastroenterology and Gastrointestinal Endoscopy, Department of Medicine, University of Verona, Piazzale L.A. Scuro, 10, 37134 Verona, Italy;* [4]*Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, United States*

Summary

Secondary analysis of large datasets involves the utilization of existing data that has typically been collected for other purposes to advance scientific knowledge. This is an established methodology applied in health services research with the unique advantage of efficiently identifying relationships between predictor and outcome variables but which has been underutilized for hepatology research. Our review of 1431 abstracts published in the 2013 European Association for the Study of Liver (EASL) abstract book showed that less than 0.5% of published abstracts utilized secondary analysis of large database methodologies.

This review paper describes existing large datasets that can be exploited for secondary analyses in liver disease research. It also suggests potential questions that could be addressed using these data warehouses and highlights the strengths and limitations of each dataset as described by authors that have previously used them. The overall goal is to bring these datasets to the attention of readers and ultimately encourage the consideration of secondary analysis of large database methodologies for the advancement of hepatology.

Introduction

Secondary analysis of large datasets involves the utilization of existing data that has typically been collected for other purposes to advance scientific knowledge. This is an established methodology applied in health services research (HSR) with the unique advantage of efficiently identifying relationships between predictor and outcome variables that could potentially serve as preliminary data for larger studies or provide hypotheses for testing using prospective study designs. Over the last three decades, the "computer revolution" has led to the generation of an increasing amount of healthcare related data mostly in the form of electronic medical records or administrative claims-based records that are stored in large databases or registries. In addition, it is now easier to survey large groups of individuals and store large quantities of data effectively. With the evolving climate in research funding, we believe that large database analyses have a greater role to play in answering pertinent research questions relating to acute and chronic liver diseases (CLD). Apart from United Network for Organ Sharing (UNOS), we posit that existing large databases are underutilized by researchers in hepatology. This is supported with evidence from our review of 1431 abstracts published in the 2013 European Association for the Study of Liver (EASL) abstract book, which showed that less than 0.5% of published abstracts utilized secondary analysis of large database methodologies. We also reviewed 2276 abstracts from the 2013 American Association for the Study of Liver Diseases (AASLD) abstract book, and found that less than 3% of published abstracts employed large database analyses methodologies. For both reviews of EASL and AASLD abstracts, we excluded UNOS abstracts.

In most cases, these large databases exist in the form of: 1) administrative and claims-based datasets; 2) clinical registries; and 3) surveys. Importantly, because observations in these datasets are occasionally generated from International Classification of Diseases (ICD-9) or Current Procedural Terminology codes, there is room for misclassification, which could significantly bias results, occasionally away from the null, leading to significant errors. To minimize misclassification with claims-based data, one strategy that could be applied involves the use of strict validation strategies in cohort identification, for instance, requiring an ICD-9 code to appear on multiple visits before inclusion. This can be combined with codes of drugs or treatments that are specific to the disease in question making the selection process more rigorous. In addition, when using hospital level datasets, the use of diagnosis related group codes may be more accurate than ICD-9 codes for identifying medical conditions.

Missing data are another limitation of large datasets. They can be missing completely at random suggesting that a missing data point is unrelated to observed and unobserved data; missing at random, suggesting that a missing data point can be explained by the observed data; or missing not at random, suggesting that

Keywords: Health services research; Liver diseases; Health care delivery research; Outcomes research.
Received 13 September 2015; received in revised form 15 December 2015; accepted 21 December 2015

* Corresponding author. Address: Division of Gastroenterology and Hepatology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, United States.
E-mail address: philokafor@post.harvard.edu (P.N. Okafor).

the missing data is dependent on unobserved values [1]. Approaches to dealing with missing data include case wise deletion where only complete cases are analyzed, use of missing indicator variables, and imputation during the process of data analysis [2]. Single value imputation methods estimate what each missing value might have been and replace it with a single value in the data set. These methods include mean imputation, last observation carried forward, and random imputation [1]. Multiple imputation methods handle missing data better by estimating and replacing missing values many times while case wise deletion or complete case analyses may introduce bias.

Another major limitation is the difficulty in establishing temporality and causality. Appropriate time must be allocated to assess the database for appropriateness to answer the research question. In most cases, the external validity of results obtained from secondary analysis of large datasets depends on the type of dataset studied. Datasets from the U.S. may have variables that are not directly applicable to patients in other countries, so results from these analyses must be critically appraised before generalization. As with any rigorous scientific study, the need for an *a priori* hypothesis before statistical analysis of these large datasets cannot be overstated. Some authors have reported that the ease of analyzing these datasets without a hypothesis often leads to their abuse, leading to results that have significant *p* values because of random effects due to the sheer large number of observations in the datasets [3], especially as some of these results may have limited clinical significance or applicability.

However, despite these limitations, compared to prospective methodologies, secondary analysis of large datasets can be performed in relatively short time periods and at relatively lower costs, usually costs associated with obtaining datasets and statistical analyses. Secondary data analysis has been employed in the description of time trends and prevalence rates of various liver diseases, and also health resource utilization. In addition, this methodology has not only been employed in the identification of variations in practice and treatment patterns of liver diseases, but also in the exploration of disparities in access to healthcare resources among patients with liver diseases, as described below. These questions would otherwise be challenging to answer using other study designs. Observations from these studies can also be used as inputs for cost-effectiveness modeling and in comparative effectiveness studies. The aim of this paper is to describe existing large databases that can be used for hepatology research. In the process, we hope to suggest potential questions that could be addressed by prospective investigators using these data warehouses. In addition, we will highlight the strengths and limitations of these datasets as described by authors that have previously used them for research. Our foremost objective is to bring these datasets to the attention of readers and ultimately encourage the consideration of secondary database methodologies in the advancement of hepatology and HSR.

Why are chronic liver diseases suitable for health services research?

CLD cause significant morbidity and mortality globally [4]. Multiple studies have demonstrated that patients with CLD (including chronic hepatitides, alcoholic and non-alcoholic fatty liver disease (NAFLD), primary biliary cirrhosis, primary sclerosing cholangitis and hepatocellular cancer (HCC)) have increased utilization of healthcare resources [5,6]. Chronic hepatitis C virus infection (HCV) remains the leading indication for liver transplant in the U.S. [7], soon to be overtaken by NAFLD. While the availability of direct-acting antivirals and implementation of HCV screening is expected to reduce the burden associated with HCV, alcoholic and NAFLD still remain significant public health problems and are increasing in prevalence [8]. With the improvement in treatment modalities for CLD, we anticipate increasing healthcare resource utilization (HCRU) by patients, which could further drive healthcare costs. However, studies to identify patient-level predictors, health system level and regional variations in HCRU, and disparities in access to healthcare among patients with liver diseases are few and far between. For this review, we will not discuss the various commercial health insurance databases which exist; instead, we will focus on some of the national databases, registries, and surveys in Europe and the U.S.

General hepatology outcomes

General Practice Research Databases (Europe)

Large general practice research databases are useful for the study of risk factors, treatment patterns, incidence rates, and health resource utilization of various medical conditions [9,10]. Among the most utilized are the General Practice Research Database (GPRD), The Health Improvement Network (THIN), and the French Longitudinal Patient Database (FLPD). The GPRD was a prospectively maintained primary care database developed in the United Kingdom (UK). It contained records of diagnoses, laboratory tests, consultations and medications from primary care and also information communicated by hospitals in discharge summaries or outpatient letters. It had previously been shown to be a representative 4–6% sample of the UK population and its accuracy and completeness have been validated in several studies. Researchers have used it for pharmacoepidemiological, pharmacoeconomics, and outcomes studies in patients with liver diseases [11–13]. In 2012, the GPRD was replaced by the Clinical Practice Research Datalink (CPRD) http://www.cprd.com/intro.asp, a new observational data and interventional research service, jointly funded by the NHS, National Institute for Health Research and the Medicines and Healthcare products Regulatory Agency [14]. The major strengths of the CPRD lie in its coverage of different patient demographics, and also in the prospective nature of data collection that minimizes recall bias. The large size affords the ability to study rare diseases. It has also been used to study a wide variety of liver diseases including drug-induced liver diseases, cirrhosis, and HCC [10,15]. Access to the CPRD comes with a price, and data costs are charged at a fixed rate depending on data sources and linkage. Data can be tailored to specific interests of the investigator. The CPRD is comparable to the FLPD, which also provides representative data from patients in France who have been treated in general practitioner facilities [16,17].

THIN is a de-identified database of patient information from over 450 general practices in the UK using the INPS vision software (http://www.thin-uk.net). It has data on over 11 million electronic patient records and has been determined to be representative of the UK population in terms of age, gender, medical conditions (including alcohol use) [18]. Like the CPRD, it contains data on diagnoses, symptoms, prescriptions, test and results. Secondary data on hospitalizations, outpatient consultations and tests are entered retrospectively. Investigators have used THIN