UROLOGIC
ONCOLOGY

Review article

Secondary data sources for health services research in urologic oncology

Alexander P. Cole, M.D., David F. Friedlander, M.D., M.P.H., Quoc-Dien Trinh, M.D.*

Division of Urological Surgery and Center for Surgery and Public Health, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

Abstract

Background: Though secondary data analyses of large datasets may reduce logistical and financial barriers required to perform significant and innovative work, such research requires specialized skills in data handling and statistical techniques as well as thorough and detailed knowledge of the data sources being used.

Objectives: To provide an overview of several common types of secondary data, focusing on strengths, weaknesses and examples of how these data may be used for health services research.

Results: Secondary data comprise a broad and heterogeneous category. This review covers several large categories of such data with examples of their use and discussions about their strengths and weaknesses. Sources include administrative data, claims-based datasets, electronic health records health surveys, patient or disease or both registries, quality improvement initiatives, as well as data from existing trials. Linkages of different types of data may expand the scope of questions answerable using secondary data analysis. Specific strengths and weaknesses of each type of dataset are discussed along with examples from the recent urologic literature.

Conclusions: Choice of the appropriate data source should be tailored to the specific research question as well as the research resources and expertise available. Appropriate decisions about which data to use are the foundation for valid, high-impact research using secondary data. © 2017 Elsevier Inc. All rights reserved.

1. Introduction

Secondary data analysis comprises a heterogeneous group of techniques that utilize data collected before defining a specific research question. In fields such as business, economics, or sociology, which do not lend themselves to prospective trials or controlled experiments, such approaches have been a mainstay for years [1]. Medical research using large secondary datasets is not new, but its popularity has grown in recent years, especially in comparative effectiveness and health services research. Compared to collecting primary data, using secondary data entails fewer logistical and financial challenges. In addition, many pressing questions in health care concern topics such as time trends, geographic variation, disparities, and outcomes of "real-world" care. Such questions may have substantial policy implications and are often better (or only) studied with large, real-world datasets [2] this is especially true for studies on costs and value in care, which represents a perennially high-impact policy questions in contemporary health research [3,4].

Despite clear benefits, there are many pitfalls for researchers seeking to answer a research question with secondary data. In addition to defining and planning appropriate analytic techniques [5], choosing the *right data* drastically impacts the success of such research. Beyond just knowing whether or not a specific variable is contained within the dataset, it is vital to understand *how* the data was collected, *why* it was collected, and how this may skew results.

In this review, we discuss some of the most common sources of secondary data for health services

Table 1
Key issues for administrative datasets

Compatibility with research question	Does the data source contain the information you need for a given question?
Availability and expense	Do I already have the dataset? Is it easy to get?
Expertise	Does my programmer know how to use this dataset?
Sample size	How frequent is the event I am studying within the dataset?
Measures of interest present and valid	For example, quality of life outcomes within a claims-based dataset
Messy data? missing data?	How often are variables missing? Have they been imputed?

research. Some of the issues we feel are most important for choosing an appropriate dataset are summarized in Table 1. We will systematically describe 7 common types of secondary data (Table 2) along with examples (mainly drawn from US datasets) below. In addition to describing some of the unique features, strengths and drawbacks of these data we will provide examples of what we believe are high-quality research studies using each type of data.

2. Types of secondary data

2.1. Administrative datasets

Administrative datasets include data collected (frequently by governments or other large health organizations) for the purposes of running or "administering" health care systems. These data are commonly collected for financial, quality improvement, monitoring, or reporting purposes but can also be used for comparative effectiveness research. Administrative data can be collected by private companies such as Premier Healthcare [6] or by one of many governmental organizations. One of the largest sets of administrative data in the United States is the Healthcare

Cost and Utilization Project (HCUP), which comprises several datasets administered by the Agency for Healthcare Research and Quality (AHRQ).

2.1.1. Healthcare Cost and Utilization Project

The HCUP (pronounced "H-Cup") is a joint project between multiple stakeholders in the United States (generally state-level organizations), which is centrally coordinated by the AHRQ. Data in HCUP generally arise from an event such as a hospital admission, readmissions, emergency room visits, etc. Hospitals generate data when one of these events occurs, which are in turn reported to various state organizations (depending on the specific state participating), which then format the data and pass them along to the centralized HCUP data aggregators.

HCUP includes a variety of datasets that collect information based off a specific "trigger" event such as hospital discharges (National Inpatient Sample, NIS), readmissions (National Readmissions Database), Emergency Department visits (State Emergency Department Databases), etc. Benefits of such data include their true "population-based" nature. Drawbacks include generally lower granularity of these data as well as the inability to longitudinally follow patients across multiple sites or episodes of care. For example, patients who get readmitted for a chronic

Table 2
Types of secondary data sources

Type of data	Characteristics	Examples
Administrative Datasets	Primarily hospital discharge data reported to a government agency	Healthcare Cost and Utilization Project (HCUP)
Claims-based datasets	Billable interactions between patients and providers from government (e.g., Medicare) or private (e.g., United Healthcare)	Medicare claims
Electronic health records	Data obtained at the point-of-care at a hospital or other health care facility	Partners health care, Kaiser
Health surveys	Often collected by phone by government agency using population-based weighting	National Health Interview Survey (NHIS)National Health and Nutrition Examination Survey (NHANES) Behavioral Risk Factor Surveillance System (BRFSS)
Patient or disease Registries	Track a narrow range of key data for certain specific conditions such as cancer	Surveillance Epidemiology and End Results (SEER) at the National Cancer InstituteNational Cancer DatabaseNational Trauma DatabankProstate Cancer Database Sweden (PcBaSe)
Quality improvement registries	Often specialty specific opt-in registries which employ self-reported data	AUA Quality Registry (AQUA)Michigan Urological Surgery Improvement Collaborative (MUSIC)
Clinical trials data	Historical, patient-level data from clinical trials	Southwest Oncology Group (SWOG)PLCO

PLCO = Prostate Lung Colorectal and Ovarian Cancer Screening Trial.