Chapter 3
Observational Studies

Jaeyong Bae and Carl V. Asche

Abstract This chapter discusses the use of secondary databases in outcomes research. Secondary databases, such as administrative databases and clinical registries, complement randomized controlled trials (RCTs) by evaluating large sample populations with a broader scope.

Secondary databases are also relatively more feasible than RCTs due to lower costs and greater timeliness. While administrative databases are originally created with the purpose of medical billing and administrative services, clinical registries are collected to assess and improve quality and outcomes of care. Administrative databases used for health outcomes research include (1) hospital inpatient discharge data, (2) ambulatory visits data, (3) emergency department visits data, and (4) health insurance claims data for both private and public insurers. Clinical registries contain more detailed clinical information on diagnoses, treatments, and prescriptions than administrative databases.

Introduction

Health outcomes research studies the effects of health-care interventions, the health-care process, and the structure of the health-care delivery system. Outcomes research is conducted by two primary means: RCTs and observation studies using secondary data.

J. Bae, PhD (✉)
Public Health and Health Education Programs, School of Nursing and Health Studies,
Northern Illinois University, DeKalb, IL, USA
e-mail: jaeyong.bae@niu.edu

C.V. Asche, PhD
Research Professor, Director of Center for Outcomes Research, Department of Pharmacy
Systems, Outcomes and Policy, Affiliate Faculty, Center for Pharmacoepidemiology and
Pharmacoeconomic Research, University of Illinois at Chicago College of Pharmacy,
Chicago, IL, USA

Research Affiliate, Centre on Aging,
University of Victoria, Victoria, British Columbia
e-mail: cva@uic.edu

C. Asche (ed.), *Applying Comparative Effectiveness Data to Medical Decision Making: A Practical Guide*, DOI 10.1007/978-3-319-23329-1_3

RCTs, widely regarded as the “gold standard” for assessment of health-care intervention, guarantee excellent internal validity because random assignment of subjects to control and treatment groups helps to ensure that the only difference between groups is their exposure to the intervention. Despite their inherent advantage in internal validity, a major limitation of RCTs lies in lack of external validity or generalizability. RCTs collect data from a targeted population with specific clinical relevance, and the sample size is typically limited.

Observational studies using secondary databases may fill the gaps of RCTs and address the lack of generalizability by assessing large sample populations with a broader scope. In addition to its advantage of external validity, the secondary data analysis is more feasible than RCTs due to lower costs and greater timeliness.

The main purpose of this chapter is to overview the use of secondary databases in outcomes research and to introduce commonly used secondary databases.

Types of Secondary Databases

Two primary types of secondary databases used in outcomes research are administrative databases and clinical registries. An administrative database is originally created with the purpose of supporting medical billing and administrative services. In contrast to administrative data, clinical registries are collected to assess and improve quality and outcomes of care. Provider and hospital characteristics data such as the American Hospital Association’s annual survey and area characteristics data such as Area Health Resources Files are also frequently used in outcomes research.

Administrative Database

Two main producers of administrative data are health insurers and government agencies. While health insurers create administrative databases such as health insurance claims data for their financial and administrative purposes, federal and state government agencies collect administrative data such as hospital discharge abstracts to track and monitor health-care utilization and outcomes. Types of administrative data used for health outcomes research include (1) hospital inpatient discharge data, (2) ambulatory visits data, (3) emergency department visits data, and (4) health insurance claims data for both private and public insurers.

Strengths and Limitations of Administrative Data

There are several advantages to using administrative data in health outcomes research. They are readily available, are feasible to obtain and analyze, and cover a large number of longitudinal cases with various clinical situations [1]. Since administrative data are already collected for their primary purposes such as medical billing and administration, they are relatively inexpensive to acquire and use. In addition, administrative databases allow for tracking patients and providers over time and assessing time trends because they are made up of population-based longitudinal data.

Limitations of using administrative databases stem from the fact that they are not collected for research purposes but for medical billing and administration. The two primary limitations are coding inaccuracy and lack of clinical detail. Records reported in administrative data lack clinical details, such as patients' comorbidity and severity of illness, for comprehensive risk adjustment and analysis. Inaccurate or incomplete coding of diagnoses and procedures in administrative data is another significant threat to internal validity.

Use of Diagnostic and Procedure Codes in Administrative Data

Administrative data contain information on diagnoses and procedures. Diagnoses and procedures are typically coded using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) classification system. ICD-9-CM diagnosis and procedure codes provide information on the condition of a patient and medical services delivered, respectively.

Comorbidities or chronic conditions of patients are used for risk adjustment in health outcomes research. However, in contrast to clinical data, administrative data do not report detailed clinical information regarding comorbidities and complications. This lack of information may be addressed by using comorbidity indexes. Two commonly used methods to identify comorbidities using administrative data are the Charlson index and the Elixhauser index [2, 3]. The Charlson Comorbidity Index contains 19 comorbidities, defined using ICD-9-CM diagnosis codes. It also provides weighed scores to predict mortality and other outcomes. As an alternative to the Charlson Comorbidity Index, Elixhauser and her colleagues developed a list of 30 comorbidities based on ICD-9-CM diagnosis and procedure codes.

In addition to comorbidities, complications or adverse events can be captured using diagnosis and procedure codes in administrative data. For example, the Agency for Healthcare Research and Quality (AHRQ) developed patient safety indicators (PSIs) to identify hospital complications and adverse events following surgeries, procedures, and childbirth [4].

Examples

Health-Care Utilization Project

The Healthcare Cost and Utilization Project (HCUP) is a set of health-care databases maintained by the AHRQ [5]. The HCUP is the largest encounter-level hospital administrative database with all-payer information in the United States.

It contains more than 100 clinical and nonclinical variables including principal and secondary diagnoses and procedures, patient demographics, admission/discharge status, total charges, length of stay, and information on the primary payer for each hospital stay.

The HCUP consists of a series of national and state-level databases: (1) the National Inpatient Sample, (2) the State Inpatient Databases, (3) the Nationwide Emergency Department Sample, (4) the State Emergency Department Databases, (5) the State Ambulatory Surgery and Services Databases, and (6) the Kids' Inpatient Database. The National Inpatient Sample (NIS) is a nationally representative database of US hospital inpatient discharges. It is a 20 % stratified sample of discharges from all US community hospitals, containing 7 million hospital stays each year. Each discharge record includes a specific weight, which enables the generation of national estimates of inpatient care utilization. The State Inpatient Databases (SID) are state-specific databases capturing all inpatient discharges for 47 participating states. The HCUP also includes other nationwide and state-specific databases of hospital-based emergency department visits (the Nationwide Emergency Department Sample and the State Emergency Department Database), ambulatory surgery and other hospital-based outpatient visits (the State Ambulatory Surgery and Services Databases), and pediatric inpatient visits (the Kids' Inpatient Database).

National Hospital Discharge Survey

The National Hospital Discharge Survey (NHDS) is a representative national sample of inpatient discharges from nonfederal short-stay hospitals, administered by the National Center for Health Statistics (NCHS) and the Centers for Disease Control and Prevention (CDC) [6, 7]. For example, the 2007 NHDS collected data for approximately 366,000 inpatient discharges from 422 hospitals, equivalent to a national estimate of 34.4 million discharges [7].

The NHDS contains information on patient demographic characteristics, principal and secondary diagnoses and procedures, length of stays, sources of payment, and patient discharge disposition. The NHDS uses multistage probability sampling procedures to allow for generating nationally representative estimates of inpatient services.

National Ambulatory Medical Care Survey and National Hospital Ambulatory Medical Care Survey

The National Ambulatory Medical Care Survey (NAMCS) and the National Hospital Ambulatory Medical Care Survey (NHAMCS) are representative national samples for ambulatory visits [8–11]. Both surveys are administered by the National Center for Health Statistics (NCHS) at the Centers for Disease Control and Prevention (CDC). NAMCS collects health-care data provided by nonfederal office-based physicians, whereas NHAMCS collects health-care data provided by nonfederal hospital outpatient departments (OPDs) and hospital emergency departments (EDs). Both NAMCS and NHAMCS contain information on patient demographic characteristics, health status (reason for visit, chronic conditions), physician and hospital characteristics (specialty, practice ownership, predominant payer model), and geographic characteristics. Both surveys use multistage probability sampling procedures to allow for generating nationally representative estimates of ambulatory medical care services in the United States.

Medicare Databases

Medicare is a federally funded program that provides health insurance coverage primarily for the elderly [12]. More than 97 % of persons aged 65 or older in the United States are covered by the Medicare program. All Medicare beneficiaries are entitled to coverage of acute care hospitalizations and stays in skilled nursing facilities (Part A), and most of the Part A beneficiaries also enroll in Part B, which covers physician, outpatient, and home health services. The Centers for Medicare & Medicaid services (CMS), which administers Medicare, collects and maintains claims records of inpatient and outpatient visits for Medicare beneficiaries.

Medicare claims include diagnosis and procedure codes as well as beneficiary, physician, and facility identifiers, which enable researchers to track health-care utilization and outcomes across multiple providers. Primarily based on claims records, CMS constructs and houses a variety of administrative databases available to health services researchers [13]. One example of these databases is the Medicare Provider Analysis and Review (MEDPAR) file containing claims records for inpatient hospital and skilled nursing facility stays for fee-for-service Medicare beneficiaries [14].

Medicaid Databases

Medicaid is a joint federal-state program that pays for health-care services for the poor and the disabled. Medicaid is administered by individual states with federal oversight by CMS. Medicaid claims data is a unique and crucial resource to assess health-care utilization and outcomes of underserved groups such as racial and ethnic minorities, low-income people, and people living with disabilities.

Medicaid databases are obtained from either CMS or individual states. CMS collects Medicaid eligibility and claims data reported by individual states through the Medicaid Statistical Information System (MSIS) [15]. Several databases, including

Medicaid Analytic eXtract (MAX) data, are extracted from the MSIS [16]. The MAX data are a set of beneficiary-level data files containing information on eligibility, demographics, and claims for inpatient care, long-term-care, and other services including ambulatory care and prescription drugs [17]. State-specific Medicaid claims data are also available through individual states.

MarketScan Databases

The Truven Health MarketScan Commercial Claims and Encounters database is a proprietary database that contains enrollment information and claims data for inpatient care, outpatient care, and drug prescriptions from large self-insured employers [18]. There are several advantages of using the MarketScan claims database in outcomes research [19]. First, the MarketScan database captures a full spectrum of care including inpatient services, ambulatory care services, and drug prescriptions. The database also enables researchers to track patients across health plans over multiple years. Finally, the MarketScan database contains comprehensive information on outpatient prescriptions to track drug use patterns and prescription trends.

Clinical Registries

Clinical registries are collected to identify patients with particular diseases and conditions, track clinical practice patterns and outcomes, and improve quality and outcomes of care. They contain more detailed clinical information on diagnoses, treatments, and prescriptions than administrative databases.

Strengths and Limitations of Clinical Registries

Clinical registries are more appropriate for outcomes research than administrative databases are because the data are collected for the purpose of quality measurement and improvement. They are also considered as the most effective resources to measure quality of care and assess effectiveness of interventions. Data derived from clinical registries have advantages in completeness and accuracy compared to administrative databases. Clinical registries also have limitations. Clinical registries are more burdensome and costly to collect than administrative databases. In addition, they are generally focused on specific conditions or populations.

Examples

Surveillance, Epidemiology, and End Results (SEER) Program

The Surveillance, Epidemiology, and End Results (SEER) Program is a population-based cancer registry maintained by the National Cancer Institute, representing 28 % of the US population [20]. The SEER Program registries contain data on patient demographics, primary tumor sites, stages of cancer, dates of cancer diagnosis, dates of death, and causes of death.

Provider Characteristic Data

American Hospital Association Annual Survey databases

The American Hospital Association (AHA) Annual Survey databases have more than 1,000 fields of data for over 6,000 hospitals representing all short-term general acute hospitals in the United States. They include demographic information, organizational structure, facilities and services, utilization data, community orientation indicators, physician arrangements, expenses, and staffing [21]. The AHA annual survey is linked to other administrative databases to provide information on hospital characteristics such as teaching status, ownership, number of beds, and rural/urban location.

American Medical Association Physician Masterfile

American Medical Association (AMA) Physician Masterfile contains current and historical information on all physicians in the United States [22]. Specifically, the AMA Physician Masterfile includes data including physician name, demographic information, address, history of prior locations, type of practice, and medical school information.

Area Characteristic Data

Area Health Resources Files

Area Health Resources Files (AHRF) data contain more than 6,000 variables for each of the nation’s counties and are available from the Health Resources and Services Administration (HRSA) to link data on health facilities, health professions, measures of resource scarcity, population health status, area economic activity, health training programs, and socioeconomic and environmental characteristics of individual counties [23].

Dartmouth Atlas of Health-Care Data

The Dartmouth Atlas of Health Care database provides a variety of measures at different geographic levels, including state, county, hospital service area (HSA), and hospital referral region (HRR) [24]. Measures in the Dartmouth Atlas database include Medicare spending and mortality rates, variation in the care of surgical conditions, and health-care supply such as number of hospital beds and physicians by specialties.

Conclusion

Observational studies using secondary databases have been prevalently used in outcomes research due to their advantages in generalizability and feasibility. There are two primary types of secondary databases: administrative databases and clinical registries.

Provider/hospital characteristics data and area characteristics data are linked with administrative data or clinical registries to provide information on provider and area characteristics. There are numerous administrative and clinical registry data sets available with their own strengths and weaknesses. Thorough assessment of potential databases, considering their strengths and weaknesses, is necessary to conduct observational studies using secondary databases.

References

1. Iezzoni LI (1997) Assessing quality using administrative data. Ann Intern Med 127(8_Part_2): 666–674
2. Charlson ME et al (1987) A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 40(5):373–383
3. Elixhauser A et al (1998) Comorbidity measures for use with administrative data. Med Care 36(1):8–27
4. Agency for Healthcare Research and Quality (2006) Patient safety indicators. [cited 28 Nov 2014]. Available from: http://www.qualityindicators.ahrq.gov/Downloads/Modules/PSI/V30/2006-Feb-PatientSafetyIndicators.pdf
5. Healthcare Cost and Utilization Project (2014) Databases. [cited 30 Nov 2014]. Available from: http://www.hcup-us.ahrq.gov/databases.jsp
6. Centers for Disease Control and Prevention (2014) National hospital discharge survey. [cited 30 Nov 2014]. Available from: http://www.cdc.gov/nchs/nhds.htm
7. Hall MJ et al (2010) National hospital discharge survey: 2007 summary. Natl Health Stat Rep 2010(29):1–20

8. Hing E et al (2010) National hospital ambulatory medical care survey: 2007 outpatient department summary. Natl Health Stat Rep 2010(28):1–32
9. Hsiao C-J et al (2010) National ambulatory medical care survey: 2007 summary. Natl Health Stat Rep 2010(27):1–32
10. Niska R, Bhuiya F, Xu J (2010) National hospital ambulatory medical care survey: 2007 emergency department summary. Natl Health Stat Rep 2010(26):1–31
11. Schappert SM, Rechtsteiner EA (2011) Ambulatory medical care utilization estimates for 2007. Vital Health Stat 13 (169):1–38
12. Centers for Medicare and Medicaid Services (2014) Medicare. [cited 30 Nov 2014]. Available from: http://www.cms.gov/Medicare/Medicare.html?redirect=/home/medicare.asp
13. Centers for medicare and Medicaid Services (2014) Research, statistics, data & systems. [cited 30 Nov 2014]. Available from: http://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems.html
14. Centers for Medicare and Medicaid Services (2014) Medicare Provider Analysis and Review (MEDPAR) file. [cited 30 Nov 2014]. Available from: http://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/IdentifiableDataFiles/MedicareProviderAnalysisandReviewFile.html
15. Centers for Medicare and Medicaid Services (2014) Medicaid statistical information system. [cited 30 Nov 2014]. Available from:http://www.medicaid.gov/medicaid-chip-program-information/by-topics/data-and-systems/msis/medicaid-statistical-information-system.html
16. Centers for Medicare and Medicaid Services (2014) Medicaid Analytic eXtract (MAX) general information. [cited 30 Nov 2014]. Available from: http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation.html
17. Borck R, Zlatinov A, Williams S (2012) The medicaid analytic eXtract 2008 Chartbook. Mathematica Policy Research
18. Truven Health Analytics (2014) Marketscan research databases. [cited 30 Nov 2014]. Available from: http://truvenhealth.com/your-healthcare-focus/analytic-research/marketscan-research-databases
19. Hansen L, Chang S (2011) Health research data for the real world: the MarketScan databases. Truven Health Analytics Inc, Ann Arbor
20. National Cancer Institute (2014) Overview of the SEER program. [cited 30 Nov 2014]. Available from: http://seer.cancer.gov/about/overview.html
21. American Hospital Association (2014) AHA annual survey database. [cited 30 Nov 2014]. Available from: http://www.ahadataviewer.com/book-cd-products/AHA-Survey/
22. American Medical Association (2014) AMA physician masterfile. [cited 30 Nov 2014]. Available from: http://www.ama-assn.org/ama/pub/about-ama/physician-data-resources/physician-masterfile.page?
23. Health Resources and Services Administration (2014) Overview: Area Health Resources Files. [cited 30 Nov 2014]. Available from: http://ahrf.hrsa.gov/overview.htm
24. Dartmouth Atlas of Health Care (2014) Atlas downloads. [cited 30 Nov 2014]. Available from: http://www.dartmouthatlas.org/tools/downloads.aspx