Chapter 8. Selection of Data Sources

Cynthia Kornegay, Ph.D.*
U.S. Food and Drug Administration, Silver Spring, MD

Jodi B. Segal, M.D., MPH
Johns Hopkins University, Baltimore, MD

Abstract

The research question dictates the type of data required, and the researcher must best match the data to the question or decide whether primary data collection is warranted. This chapter discusses considerations for data source selection for comparative effectiveness research (CER). Important considerations for choosing data include whether or not the key variables are available to appropriately define an analytic cohort and identify exposures, outcomes, covariates, and confounders. Data should be sufficiently granular, contain historical information to determine baseline covariates, and represent an adequate duration of followup. The widespread availability of existing data from electronic health records, personal health records, and drug surveillance programs provides an opportunity for answering CER questions without the high expense often associated with primary data collection. If key data elements are unobtainable in an otherwise ideal dataset, methods such as predicting absent variables with available data or interpolating for missing time points may be used. Alternatively, the researcher may link datasets. The process of data linking, which combines information about one individual from multiple sources, increases the richness of information available in a study. This is in contrast to data pooling and networking, which are normally used to increase the size of an observational study. Each data source has advantages and disadvantages, which should be considered thoroughly in light of the research question of interest, as the validity of the study will be dictated by the quality of the data. This chapter concludes with a checklist of key considerations for selecting a data source for a CER protocol.

Introduction

Identifying appropriate data sources to answer comparative effectiveness research (CER) questions is challenging. While the widespread availability of existing data provides an opportunity for answering CER questions without the high expense associated with primary data collection, the data source must be chosen carefully to ensure that it can address the study question, that it has a sufficient number of observations, that key variables are available, that there is adequate confounder control, and that there is a sufficient length of followup.

This chapter describes data that may be useful for observational CER studies and the sources of these data, including data collected for both research and nonresearch purposes. The chapter also explains how the research question should dictate the type of data required and how to best match data to the issue at hand. Considerations for evaluating data quality (e.g., demonstrating data integrity) and privacy protection provisions are discussed. The chapter concludes by describing new sources of data that may expand the options available to CER researchers to address questions. Recommendations for "best practices" regarding data selection are included, along with a checklist that researchers may use when developing and writing a CER protocol. To start, however, it is important to consider primary data collection for observational research, since the use of secondary data may be impossible or unwise in some situations.

**Disclaimer:* The views expressed are the authors' and not necessarily those of the Food and Drug Administration.

Data Options

Primary data are data collected expressly for research. Observational studies, meaning studies with no dictated intervention, require the collection of new data if there are no adequate existing data for testing hypotheses. In contrast, secondary data refer to data that were collected for other purposes and are being used *secondarily* to answer a research question. There are other ways to categorize data, but this classification is useful because the types of information collected for research differ markedly from the types of information collected for nonresearch purposes.

Primary Data

Primary data are collected by the investigator directly from study participants to address a specific question or hypothesis. Data can be collected by in-person or telephone interviews, mail surveys, or computerized questionnaires. While primary data collection has the advantage of being able to address a specific study question, it is often time consuming and expensive. The observational research designs that often *require* primary data collection are described here. While these designs may also incorporate existing data, we describe them here in the context of primary data collection. The need to use these designs is determined by the research question; if the research question clearly must be answered with these designs below, primary data collection may be required. Additional detail about the selection of suitable study design for observational CER is presented in chapter 2.

Prospective Observational Studies

Observational studies are those in which individuals are selected on the basis of specific characteristics and their progress is monitored. A key concept is that the investigator does not assign the exposure(s) of interest. There are two basic observational designs: (1) cohort studies, in which selection is based on exposure and participants are followed for the occurrence of a particular outcome, and (2) case-control studies, where selection is based on a disease or condition and participants are contacted to determine a particular exposure.

Within this framework, there is a wide variety of possible designs. Participants can be individuals or groups (e.g., schools or hospitals); they can be followed into the future (prospective data collection) or asked to recall past events (retrospective data collection); and, depending on the specific study questions, elements of the two basic designs can be combined into a single study (e.g., case-cohort or nested case-control studies). If information is also collected on those who are either not exposed or do not have the outcome of interest, observational studies can be used for hypothesis testing.

An example of a prospective observational study is a recent investigation comparing medication adherence and viral suppression between once-daily and more-than-once daily pill regimens in a homeless and near-homeless HIV-positive population.[1] Adherence was measured using unscheduled pill-count visits over the six-month study period while viral suppression was determined at the end of the study. The investigators found that both adherence and viral suppression levels were higher in the once-daily groups compared to the more-than-once-daily groups. The results of this study are notable as they indicate an effective method to treat HIV in a particularly hard-to-reach population.

Registries

In the most general sense, a registry is a systematic collection of data. Registries that are used for research have clearly stated purposes and targeted data collection.

Registries use an observational study design that does not specify treatments or require therapies intended to change patient outcomes. There are generally few inclusion and exclusion criteria to make the results broadly generalizable. Patients are typically identified when they present for care, and the data collected generally include clinical and laboratory tests and measurements. Registries can be defined by specific diseases or conditions (e.g., cancer, birth defects, or rheumatoid arthritis), exposures (e.g., to drug products, medical devices, environmental conditions, or radiation), time periods, or populations. Depending on their purpose and the information collected, registry data can potentially be used for public health

surveillance, to determine incidence rates, to perform risk assessment, to monitor progress, and to improve clinical practice. Registries can also provide a unique perspective into specialized subpopulations. However, like any long-term study, they can be very expensive to maintain due to the effort required to remain in contact with the participants over extended periods of time.

Registries have been used extensively for CER. As an example, the United States Renal Data System (USRDS) is a registry of individuals receiving dialysis that includes clinical data as well as medical claims. This registry has been used to answer questions about the comparative effectiveness and safety of erythropoiesis-stimulating agents and iron in this patient population,[2] the comparative effectiveness of dialysis chain facilities,[3] and the effectiveness of nocturnal versus daytime dialysis.[4] Another registry is the Surveillance, Epidemiology, and End Results (SEER) registry, which gathers data on Americans with cancer. Much of the SEER registry's value for CER comes from its linkage to Medicare data. Examples of CER studies that make use of this linked data include an evaluation of the effectiveness of radiofrequency ablation for hepatocellular carcinoma compared to resection or no treatment[5] and a comparison of the safety of open versus radical nephrectomy in individuals with kidney cancer.[6] A third example is a study that used SEER data to evaluate survival among individuals with bladder cancer who underwent early radical cystectomy compared to those patients who did not.[7]

Secondary Data

Much secondary data that are used for CER can be considered byproducts of clinical care. The framework developed by Schneeweiss and Avorn is a useful structure with which to consider the secondary sources of data generated within this context.[8] They described the "record generation process," which is the information generated during patient care. Within this framework, data are generated in the creation of the paper-based or electronic medical (health) record, claims are generated so that providers are paid for their services, and claims and dispensing records are generated at the pharmacy at the time of payment. As data are not collected specifically for the research question of interest, particular attention must be paid to ensure that data quality is sufficient for the study purpose.

A thorough understanding of the health system in which patients receive care and the insurance products they use is needed for a clear understanding of whether the data are likely to be complete or unavailable for the population of interest. Integrated health delivery systems such as Kaiser Permanente, in which patients receive the majority of their care from providers and facilities within the system, provide the most complete picture of patient medical care.

Electronic Health Record (EHR) Data

Electronic health records (EHRs) are used by health care providers to capture the details of the clinical encounter. They are chiefly clinical documentation systems. They are populated with some combination of free text describing findings from the history and the physical examination; results inputted with check-boxes to indicate positive responses; patient-reported responses to questions for recording symptoms or for screening; prefilled templates that describe normal and abnormal findings; imported text from earlier notes on the patient; and linkages to laboratory results, radiology reports and images; and specialized testing results (such as electrocardiograms, echocardiograms, or pulmonary function test results). Some EHRs include other features, such as flow sheets of clinical results, particularly those results used in inpatient settings (e.g., blood pressure measurements); problem and habits lists, electronic medication administration records; medication reconciliation features; decision support systems and/or clinical pathways and protocols; and specialty features for the documentation needs of specialty practices. The variables that *might* be accessible from EHR data are shown in Table 8.1.

Table 8.1. Data elements available in electronic health records and/or in administrative claims data

Information	EHRs	Administrative Claims
Prescriptions ordered	Yes	No
Pharmacy data (drugs dispensed)	Sometimes	Yes
Medication list	Often	No
Inpatient medications ordered or administered	Sometimes	No
Clinical data: vital signs or point of care testing results	Yes*	No
Clinical data: inpatient	Sometimes*	No
Clinical data: outpatient	Yes*	No
Age/sex	Yes	Yes
Race/ethnicity	Sometimes	Sometimes
Socioeconomic data	Sometimes	Inferred (from zip code)
Insurance information	Yes	Yes
Spontaneously reported adverse events	Yes	Yes
Diagnoses or procedures coded for payment	No	Yes
Behavioral risk factors	Yes*	No
Diet	Sometimes*	No
Indicators of procedures having being done (laboratory, radiologic, therapeutic)	Yes	Yes
Results from diagnostic procedures (echocardiography, radiology)	Yes	No
Laboratory results	Yes	No
Problem list or summary	Yes	No

*It should be noted that clinical data available in EHRs are often missing informatively in high proportions. For example, a study examining data quality issues in an EHR-based survival analysis of patients with pancreatic cancer found that patients with late-stage ductal adenocarcinomas were more likely to have missing biochemistry lab data compared to early-stage patients (6-9% incomplete in early-stage patients versus 13-23% incomplete in late-stage patients).[9] The authors conclude that this was likely due to terminally ill patients receiving care outside of the EHR system in dedicated cancer treatment centers.

As can be seen from the variable list, the details about an individual patient may be extensive. The method of data collection is not standardized and the intervals between visits vary for every patient in accordance with usual medical practice. Of note, medication information captured in EHRs differs from data captured by pharmacy claims. While pharmacy claims contain information on medications dispensed (including the national drug code [NDC] to identify the medication, dispensing date, days' supply, and amount dispensed), EHRs more typically contain information on medications prescribed by a clinician. Medication data from EHRs are often captured as part of the patient's medication list, which may include the medication name, order date, strength, units, quantity, and frequency. Again depending on the specific EHR system, inpatient medication orders may or may not be available but are not typically. As EHRs differ substantially, it is important to understand what fields are captured in the EHR under consideration, and to realize that completeness of specific fields may vary depending on how individual health care providers use the EHR.

An additional challenge with EHR data is that patients may receive care at different facilities, and information regarding their health may be entered separately into multiple systems that are not integrated and are inconsistent across practices. If a patient has an emergency room visit at a hospital that is not his usual site of care, it is unlikely to be recorded in the electronic medical record that houses the majority of his clinical information. Additionally, for a patient who resides in two or more cities during the year, the electronic medical record at each institution may be incomplete if the institutions do not share a common data system.

Paper-Based Records

Although time-intensive to access, the use of paper-based records is sometimes required. Many practices still do not have EHRs; in 2009, it was estimated that only half of outpatient practices in the U.S. were using EHRs.[10] Exclusion of sites without electronic records may bias study results because these sites may have different patient populations or because there may be regional differences in practice. These data may be particularly valuable if patient-reported information is needed (such as severity of pain, quality of symptoms, mental health concerns, and habits). The richness of information in paper-based records may exceed that in EHR data particularly if the electronic data is template driven. Additionally, paper-based records are valuable as a source of primary data for validating data that is available elsewhere such as in administrative claims. With a paper medical record, the researcher can test the sensitivity and specificity of the information contained in claims data by reviewing the paper record to see if the diagnosis or procedure was described. In that situation, the paper-based record would be considered the reference standard for diagnoses and procedures.

Administrative Data

Administrative health insurance data are typically generated as part of the process of obtaining insurance reimbursement. Presently, medical claims are most often coded using the International Classification of Disease (ICD) and the Common Procedural Terminology (CPT) systems. The ICD, Ninth Revision, Clinical Modification (ICD-9-CM) is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States. Much of Europe is using ICD-10 already, while the United States currently uses ICD-9 for everything except mortality data; the United States will start using ICD-10 in October 2013.[11] The ICD coding terminology includes a numerical list of codes identifying diseases, as well as a classification system for surgical, diagnostic, and therapeutic procedures. The National Center for Health Statistics and the Centers for Medicare and Medicaid Services (CMS) are responsible for overseeing modifications to the ICD. For outpatient encounters, the CPT is used for submitting claims for services. This terminology was initially developed by the American Medical Association in 1966 to encourage the use of standard terms and descriptors to document procedures in the medical record, to communicate accurate information on procedures and services to agencies concerned with insurance claims, to provide the basis for a computer-oriented system to evaluate operative procedures, and for actuarial and statistical purposes. Presently, this system of terminology is the required nomenclature to report outpatient medical procedures and services to U.S. public and private health insurance programs, as the

ICD is the required system for diagnosis codes and inpatient hospital services.[12] The diagnosis-related group (DRG) classification is a system to classify hospital cases by their ICD codes into one of approximately 500 groups expected to have similar hospital resource use; it was developed for Medicare as part of the prospective payment system. The DRG system can be used for research as well, but with the recognition that there may be clinical heterogeneity within a DRG. There is no correlate of the DRG for outpatient care.

When using these claims for research purposes, the validity of the coding is of the highest importance. This is described in more detail below. The validity of codes for procedures exceeds the validity of diagnostic codes, as procedural billing is more closely tied to reimbursement. Understandably, the motivation for coding procedures correctly is high. For diagnosis codes, however, a diagnosis that is under evaluation (e.g., a medical visit or a test to "rule out" a particular condition) is indistinguishable from a diagnosis that has been confirmed. Consequently, researchers tend to look for sequences of diagnoses, or diagnoses followed by treatments appropriate for those diagnoses, in order to identify conditions of interest. Although Medicare requires an appropriate diagnosis code to accompany the procedure code to authorize payment, other insurers have looser requirements. There are few external motivators to code diagnoses with high precision, so the validity of these codes requires an understanding of the health insurance system's approach to documentation.[13-20] Investigators using claims data for CER should validate the key diagnostic and procedure codes in the study. There are many examples of validation studies in the literature upon which to pattern such a study.[18, 21-22] Additional codes are available in some datasets—for example, the "present on admission" code that has been required for Medicare and Medicaid billing since October 2007—which may help in further refinement of algorithms for identifying key exposures and outcomes.

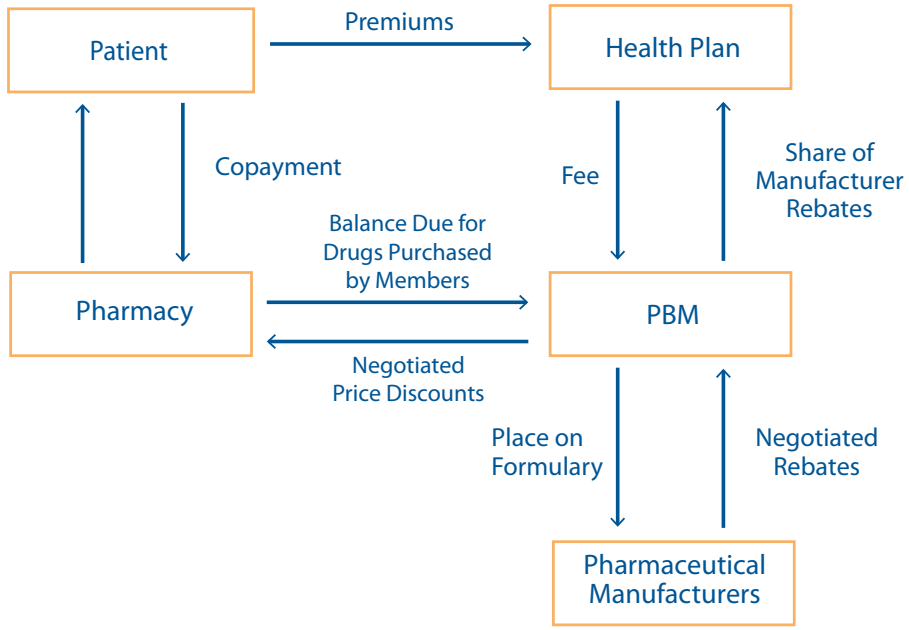
Pharmacy Data

Outpatient pharmacy data include claims submitted to insurance companies for payment as well as the records on drug dispensing kept by the pharmacy or by the pharmacy benefits manager (PBM). Claims submitted to the insurance company use the NDC as the identifier of the product. The NDC is a unique, 10-digit, 3-segment number that is a standard product identifier for human drugs in the United States. Included in this number are the active ingredient, the dosage form and route of administration, the strength of the product, and the package size and type. The U.S. Food and Drug Administration (FDA) has authority over the NDC codes. Claims submitted to insurance companies for payment for drugs are submitted with the NDC code as well as information about the supply dispensed (e.g., how many days the prescription is expected to cover), and the amount of medication dispensed. This information can be used to provide a detailed picture of the medications dispensed to the patient. Medications for which a claim is not submitted or is not covered by the insurance plan (e.g., over-the-counter medications) are not available. It should be noted that claims data are generally weak for medical devices, due to a lack of uniform coding, and claims often do not include drugs that are not dispensed through the pharmacy (e.g., injections administered in a clinic).

Large national PBMs, such as Medco Health Solutions or Caremark, administer prescription drug programs and are responsible for processing and paying prescription drug claims. They are the interface between the pharmacies and the payers, though some larger health insurers manage their own pharmacy data. PBM models differ substantially, but most maintain formularies, contract with pharmacies, and negotiate prices with drug manufacturers. The differences in formularies across PBMs may offer researchers the advantage of natural experiments, as some patients will not be dispensed a particular medication even when indicated, while other patients will be dispensed the medication, solely due to the formulary differences of their PBMs. Some PBMs own their own mail-order pharmacies, eliminating the local pharmacies' role in distributing medications. PBMs more recently have taken on roles of disease management and outcomes reporting, which generates additional data that may be accessible for research purposes. Figure 8.1 illustrates the flow of information into PBMs from health plans, pharmaceutical

manufacturers, and pharmacies. PBMs contain a potentially rich source of data for CER, provided that these data can be linked with outcomes. Examples of CERs that have been done using PBM data include two studies that evaluate patient adherence to medications as their outcome. One compared adherence to different antihypertensive medications using data from Medco Health Solutions. The researchers identified differential adherence to antihypertensive drugs, which has implications for their effectiveness in practice.[23] Another study compared costs associated with a step-therapy intervention that controlled access to angiotensin-receptor blockers with the costs associated with open access to these drugs.[24] Data came from three health plans that contracted with one PBM and one health plan that contracted with a different PBM.

Figure 8.1. How pharmacy benefits managers fit within the payment system for prescription drugs



From the Congressional Budget Office, based in part on General Accounting Office, Pharmacy Benefit Managers: Early Results on Ventures with Drug Manufacturers. GAO/HEHS-96-45. November 1995.

Frequently, PBM data are accessible through health insurers along with related medical claims, thus enabling single-source access to data on both treatment and outcomes. Data from the U.S. Department of Veterans Affairs (VA) Pharmacy Benefits Manager, combined with other VA data or linked to Medicare claims, are a valuable resource that has generated comparative effectiveness and safety information.[25-26]

Regulatory Data

FDA has a vast store of data from submissions for regulatory approval from manufacturers. While the majority of the submissions are not in a format that is usable for research (e.g., paper-based submissions or PDFs), increasingly the submissions are in formats where the data may be used for purposes beyond that for which they were collected, including CER. Additionally, FDA is committed to converting many of its older datasets into research-appropriate data. FDA presently has a contractor working on conversion of 101 trials into useable data that will be stored in their clinical trial repository.[27] It also has pilot projects underway that are exploring the benefits and risks of providing external researchers access to their data for CER. It is recognized that issues of using proprietary data or trade-secret data will arise, and that there may be regulatory and data-security challenges to address. A limitation of using these

trials for CER is that they are typically efficacy trials rather than effectiveness trials. However, when combined using techniques of meta-analysis, they may provide a comprehensive picture of a drug's efficacy and short-term safety.

Repurposed Trial Data or Data From Completed Observational Studies

A vast amount of data is collected for clinical research in studies funded by the Federal government. By law, these data must be made available upon request to other researchers, as this was information collected with taxpayer dollars. This is an exceptional source of existing data. To illustrate, the Cardiovascular Health Study is a large cohort study that was designed to identify risk factors for coronary heart disease and stroke by means of a population-based longitudinal cohort study.[28] The study investigators collected diverse outcomes including information on hospitalization, specifically heart failure associated hospitalizations. Thus, the data from this study can be used to answer comparative effectiveness questions about interventions and their effectiveness on preventing heart failure complications, even though this was not a primary aim of the original cohort study. A limitation is that the researcher is limited to only the data that were collected—an important consideration when selecting a dataset. Some of the datasets have associated biospecimen repositories from which specimens can be requested for additional testing.

Completed studies with publicly available datasets often can be identified through the National Institutes of Health institute that funded the study. For example, the National Heart Lung and Blood Institute has a searchable site (at https://biolincc.nhlbi.nih.gov/home/) where datasets can be identified and requested. Similarly, the National Institute of Diabetes and Digestive and Kidney Diseases has a repository of datasets as well as instructions for requesting data (at https://www.niddkrepository.org/niddk/jsp/public/resource.jsp).

Considerations for Selecting Data

Required Data Elements

The research question must drive the choice of data. Frequently, however, as the question is developed, it becomes clear that a particular piece of information is critical to answering the question. For example, a question about interventions that reduce the amount of albuminuria will almost certainly require access to laboratory data that include measurement of this outcome. Reliance on ICD-9 codes or use of a statement in the medical record that "albuminuria decreased" will be insufficiently specific for research purposes. Similarly, a study question about racial differences in outcomes from coronary interventions requires data that include documentation of race; this requirement precludes use of most administrative data from private insurers that do not collect this information. If the relevant data are not available in an existing data source, this may be an indication that primary data collection or linking of datasets is in order. It is recommended that the investigator specify a priori what the minimum requirements of the data are before the data are identified, as this will help avoid the effort of making suboptimal data work for a given study question.

If some key data elements seem to be unobtainable in an otherwise suitable dataset, one might consider ways to supplement the available data. These strategies may be methodological, such as predicting absent data variables with data that are available, or interpolating for missing time points. The authors recently completed a study in which the presence of obesity was predicted for individuals in the dataset based on ICD-9 codes.[29] In such instances, it is desirable to provide a reference to support the quality of data obtained by such an approach.

Alternatively, there may be a need to link datasets or to use already linked datasets. SEER-Medicare is an example of an already linked dataset that combines the richness of the SEER cancer diagnosis data with claims data from Medicare.[30] Unique patient identifiers that can be linked across datasets (such as Social Security numbers) provide opportunities for powerful linkages with other datasets.[31] Other methods have been developed that do not rely on the existence of unique identifiers.[32] As described above, linking medical data with environmental data, population-level data, or census data provides rich datasets for addressing research questions. Privacy concerns raised by individual contributors can greatly increase the complexity and time needed for a study with linked data.

Data *linking* combines information on the same person from multiple sources to increase the richness of information available in a study. This is in contrast to data *pooling* and *networking*, tools primarily used to increase the size of an observational study.

Time Period and Duration of Followup

In an ideal situation, researchers have easy access to low-cost, clinically rich data about patients who have been continuously observed for long periods of time. This is seldom the case. Often, the question being addressed is sensitive to the time the data were collected. If the question is about a newly available drug or device, it will be essential that the data capture the time period of relevance. Other questions are less sensitive to secular changes; in these cases, older data may be acceptable.

Inadequate length of followup for individuals is often the key time element that makes data unusable. How long is necessary depends on the research question; in most cases, information about outcomes associated with specific exposures requires a period of followup that takes the natural history of the outcomes into account. Data from registries or from clinical care may be ideal for studies requiring long followup. Commercial insurers see large amounts of turnover in their covered patient populations, which often makes the length of time that data are available on a given individual relatively short. This is also the case with Medicaid data. The populations in data from commercial insurers or Medicaid, however, are so large that reasonable numbers of relevant individuals with long followup can often be identified. It should be noted that when a study population is restricted to patients with longer than typical periods of followup within a database, the representativeness of those patients should be assessed. Individuals insured by Medicare are typically insured by Medicare for the rest of their lives, so these data are often appropriate for longitudinal research, especially when they can be coupled with data on drug use. Similarly, the VA health system is often a source of data for CER because of the relatively stable population that is served and the detail of the clinical information captured in the system's electronic records.

Table 8.2 provides the types of questions, with an example for each, that an investigator should ask when choosing data.

Table 8.2. Questions to consider when choosing data

Question To Ask	Example
Are the key variables available to define an analytic cohort (the study inclusion and exclusion criteria)?	Do the data contain height and weight or BMI to define a cohort of overweight or obese subjects?
Are the key variables available for identifying important subpopulations for the study?	Do the data contain a variable describing race for a study of racial differences in outcomes of coronary stenting?
Are the key variables available for identifying the relevant exposures, outcomes, and important covariates and confounders?	Do the data contain information on disease severity to assess the comparative effectiveness of conservative versus intensive management of prostate cancer? (Disease severity is a likely confounder.)
Are the data sufficiently granular for the purpose of the study?	Is it adequate to know whether the individual has hypertension or not, or is it important to know that the individual has Stage I or Stage III hypertension?
Are there a sufficient number of exposed individuals in the dataset?	Are there enough individuals who filled prescriptions for exenatide to study the outcomes from this medication?

Table 8.2. Questions to consider when choosing data (continued)

Question To Ask	Example
Do the data contain a sufficiently long duration of followup after exposures?	Are there data on weight for at least three years after bariatric surgery?
Are there sufficient historical data to determine baseline covariates?	Is there information on hospitalizations in the year prior to cardiac resynchronization therapy for an observational study of outcomes from the device?
Is there a complete dataset from all appropriate settings of care to comprehensively identify exposures and outcomes?	Is there a record of emergency department visits in addition to a record of outpatient and hospitalized care in a study of children with asthma?
Are data available on other exposures outside of the healthcare setting?	Are there data on aspirin exposure when purchased over the counter in a study of outcomes after myocardial infarction?
Are there a sufficient number of observations in the dataset if restricting the patient population is necessary for internal validity (e.g., restriction to new users)?	Are there a sufficient number of new users (based on a "washout period" of at least 6 months) of each selective and non-selective nonsteroidal anti-inflammatory drug (NSAID) to study outcomes in users of each of these medications?
What is the difference between the study and target population demographics and distributions of comorbid illnesses? Will these differences affect the interpretation and generalizability of the results?	Is the age range of the data source appropriate to address the study question? Can any differences in demographics between data source and target population be addressed through appropriate design or analysis approaches?

Ensuring Quality Data

When considering potential data resources for a study, an important element is the quality of the information in the resource. Using databases with large amounts of missing information, or that do not have rigorous and standardized data editing, cleaning, and processing procedures increases the risk of inconclusive and potentially invalid study results.

Missing Data

One of the biggest concerns in any investigation is missing data. Depending on the elements and if there is a pattern in the type and extent of missingness, missing data can compromise the validity of the resource and any studies that are done using that information. It is important to understand what variables are more or less likely to be missing, to define a priori an acceptable percent of missing data for key data elements required for analysis, and to be aware of the efforts an organization takes to minimize the amount of missing information. For example, data resources that obtain data from medical or insurance claims will generally have higher completion rates for data elements used in reimbursement, while optional items will be completed less frequently. A data resource may also have different standards for individual versus group-level examination. For example, while ethnicity might be the only missing variable in an individual record, it could be absent for a significant percentage of the study population.

Some investigators impute missing data elements under certain circumstances. For example, in a longitudinal resource, data that were previously present may be carried forward if the latest update of a patient's information is missing. Statistical imputation techniques may be used to estimate or approximate missing data by modeling the characteristics of cases with missing data to those who have such data.[33-35] Data that have been generated in this manner should be clearly identified so that they can be removed

for sensitivity analyses, as may be appropriate. Additional information about methods for handling missing data in analysis is covered in chapter 10.

Changes That May Alter Data Availability and Consistency Over Time

Any data resource that collects information over time is likely to eventually encounter changes in the data that will affect longitudinal analyses. These changes could be either a singular event or a gradual shift in the data and can be triggered by the organization that maintains the database or by events beyond the control of that organization including adjustments in diagnostic practices, coding and reimbursement modifications, or increased disease awareness. Investigators should be aware of these changes as they may have a substantial effect on the study design, time period, and execution of the project.

Sudden changes in the database may be dealt with by using trend breaks. These are points in time where the database is discontinuous, and analyses that cross over these points will need to be interpreted with care. Examples of trend break events might be major database upgrades and/or redesigns or changes in data suppliers. Other trend break events that are outside the influence of the maintenance organization might be medical coding upgrades (e.g., ICD-9 to ICD-10), announcements or presentations at conferences (e.g., Women's Health Initiative findings) that may lead to changes in medical practice, or high profile drug approvals or withdrawals.

More gradual events can also affect the data availability. Software upgrades and changes might result in more data being available for recently added participants versus individuals who were captured in prior versions. Changes in reimbursement and recommended practice could lead to shifts in use of ICD-9 codes, or to more or less information being entered for individuals.

Validity of Key Data Definitions

Validity assessment of key data in an investigation is an important but sometimes overlooked issue in health care research using secondary data. There is a need to assess not only the general definition of key variables, but also their reliability and validity in the particular database chosen for the analysis. In some cases, particularly for data resources commonly used for research, other researchers or the organization may have validated outcomes of health events (e.g., heart attack, hospitalization, or mortality).[36] Creating the best definitions for key variables may require the involvement of knowledgeable clinicians who might suggest that the occurrence of a specific procedure or a prescription would strengthen the specificity of a diagnosis. Knowing the validity of other key variables, such as race/ethnicity, within a specific dataset is essential, particularly if results will be described in these subgroups.

Ideally, validity is examined by comparing study data to additional or alternative records that represent a "gold standard," such as paper-based medical records. We described in the Administrative Data section above how validity of diagnoses associated with administrative claims might be assessed relative to paper-based records. EHRs and non–claims-based resources do not always allow for this type of assessment, but a more accommodating validation process has not yet been developed. When a patient's primary health care record is electronic, there may not be a paper trail to follow. Commonly, all activity is integrated into one record, so there is no additional documentation. On the other hand, if the data resource pulls information from a switch company (an organization that specializes in routing claims between the point of service and an insurance company), there may be no mechanism to find additional medical information for patients. In those cases, the information included in the database is all that is available to researchers.

Data Privacy Issues

Data privacy is an ongoing concern in the field of health care research. Most researchers are familiar with the Health Insurance Portability and Accountability Act (HIPAA), enacted in 1996 in part to standardize the security and privacy of health care information. HIPAA coined the term "protected health information" (PHI), defined as any individually identifiable health information (45 CFR 160.103). HIPAA requires that patients be informed of the use of their PHI and that covered

entities (generally, health care clearinghouses, employer-sponsored health plans, health insurers, and medical service providers) track the use of PHI. HIPAA also provides a mechanism for patients to report when they feel these regulations have been violated.[37]

In practical terms, this has resulted in an increase in the amount and complexity of documentation and permissions required to conduct healthcare research and a decrease in patient recruitment and participation levels.[38-39] While many data resources have established procedures that allow for access to data without personal identifiers, obtaining permission to use identifiable information from existing data sources (e.g., from chart review) or for primary data collection can be time consuming. Additionally, some organizations will not permit research to proceed beyond a certain point (e.g., beginning or completing statistical analyses, dissemination, or publication of results) without proper institutional review board approvals in place. If a non-U.S. data resource is being used, researchers will need to be aware of differences between U.S. privacy regulations and those in the country where the data resource resides.

Adherence to HIPAA regulations can also affect study design considerations. For example, since birth, admission, and discharge dates are all considered to be PHI, researchers may need to use a patient's age at admission and length of stay as unique identifiers. Alternatively, a limited data set that includes PHI but no direct patient identifiers such as name, address, or medical record numbers may be defined and transferred with appropriate data use agreements in place. Organizations may have their own unique limits on data sharing and pooling. For example, in the VA system, the general records and records for condition-specific treatment, such as HIV treatment, may not be pooled. Additional information regarding HIPAA regulations as they apply to data used for research may be found on the National Institutes of Health Web site.[40]

Emerging Issues and Opportunities

Data From Outside of the United States

Where appropriate, non-U.S. databases may be considered to address CER questions, particularly for longitudinal studies. One of the main reasons is that, unlike the majority of U.S. health care systems, several countries with single-payer systems, such as Canada, the United Kingdom, and the Netherlands, have regional or national EMR systems. This makes it much easier to obtain complete, long-term medical records and to follow individuals in longitudinal studies.[41]

The Clinical Practice Research Datalink (CPRD) is a collection of anonymized primary care medical records from selected general practices across the United Kingdom. These data have been linked to many other datasets to address comparative effectiveness questions. An example is a study that linked the CPRD to the Myocardial Ischaemia National Audit Project registry in England and Wales. The researchers answered questions about the risks associated with discontinuing clopidogrel therapy after a myocardial infarction (performed when the database was called General Practice Research Database).[42]

While the selection of a non-U.S. data source may be the right choice for a given study, there are a number of things to consider when designing a study using one of these resources.

One of the main considerations is if the study question can be appropriately addressed using a non-U.S. resource. Questions that should be addressed during the study design process include:

- Is the exposure of interest similar between the study and target population? For example, if the exposure is a drug product, is it available in the same dose and form in the data resource? Is it used in the same manner and frequency as in the United States?

- Are there any differences in availability, cost, practice, or prescribing guidelines between the study and target populations? Has the product been available in the study population and the United States for similar periods of time?
- What is the difference between the health care systems of the study and target populations? Are there differences in diagnosis methods and treatment patterns for the outcome of interest? Does the outcome of interest occur with the same frequency and severity in the study and target populations?
- Are the comparator treatments similar to those that would be available and used in the United States?

An additional consideration is data access. Access to some resources, such as the United Kingdom's CPRD, can be purchased by interested researchers. Others, such as Canada's regional health care resources, may require the personal interest of and an official association with investigators in that country who are authorized to use the system. If a non-U.S. data resource is appropriate for a proposed study, the researcher will need to become familiar with the process for accessing the data and allow for any extra time and effort required to obtain permission to use it.

A sound justification for selecting a non-U.S. data resource, a solid understanding of the similarities and differences of the non-U.S. versus the U.S. systems, as well as careful discussion of whether the results of the study can be generalized to U.S. populations will help other researchers and health care practitioners interpret and apply the results of non-U.S.-based research to their particular situations.

Point of Care Data Collection and Interactive Voice Response/Other Technologies

Traditionally, the data used in epidemiologic studies have been gathered at one point in time, cleaned, edited, and formatted for research use at a later point. As technology has developed, however, data collected close to the point of care increasingly have been available for analysis. Prescription claims can be available for research in as little as one week.

In conjunction with a shortened turnaround time for data availability, the point at which data are coded and edited for research is also occurring closer to when the patient received care. Many people are familiar with health care encounters where the physician takes notes, which are then transcribed and coded for use. With the advent of EHRs, health information is now coded and transcribed into a searchable format at the time of the visit; that is, the information is directly coded as it is collected, rather than being transcribed later.

Another innovation is using computers to collect data. Computer-aided data collection has been used in national surveys since the 1990s[43] and also in types of research (such as risky behaviors, addiction, and mental health) where respondents might not be comfortable responding to a personal interviewer.[44-46]

The advantages of these new and timely data streams are more detailed data, sometimes available in real or near-real time that can be used to spot trends or patterns. Since data can be recorded at the time of care by the health care provider, this may help minimize miscoding and misinterpretation. Computerized data collection and Interactive Voice Response are becoming easier and less expensive to use, enabling investigators to reach more participants more easily. Some disadvantages are that these data streams are often specialized (e.g., bedside prescribing), and, without linkage to other patient characteristics, it can be difficult to track unique patients. Also, depending on the survey population, it can be challenging to maintain current telephone numbers.[47-48]

Data Pooling and Networking

A major challenge in health research is studying rare outcomes, particularly in association with common exposures. Two methods that can be used to address this challenge are data pooling and networking. Data pooling is combining data, at the level of the unit of analysis (i.e., individual), from several sources into a single cohort for analysis. Pooled data may also include data from unanalyzed and unpublished investigations, helping to minimize the potential for publication bias. However, pooled analyses require close

coordination and can be very difficult to complete due to differences in study methodology and collection practices. An example is an analysis that pooled primary data from four cohorts of breast cancer survivors to ask a new question about the effectiveness of physical activity. The researchers had to ensure the comparability of the definitions of physical activity and its intensity in each cohort.[49] Another example is a study that pooled data from four different data systems including from Medicare, Medicaid, and a private insurer to assess the comparative safety of biological products in rheumatologic diseases. The authors describe their assessment of the comparability of covariates across the data systems.[50] Researchers must be sensitive to whether additional informed consent of individuals is needed for using their data in combination with other data. Furthermore, privacy concerns sometimes do not allow for the actual combination of raw study data.[51]

An alternative to data pooling is data networking, sometimes referred to as virtual data networks or distributed research networks. These networks have become possible as technology has developed to allow more sophisticated linkages. In this situation, common protocols, data definitions, and programming are developed for several data resources. The results of these analyses are combined in a central location, but individual study data do not leave the original data resource site. The advantage of this is that data security concerns may be fewer. As with data pooling, the differences in definitions and use of terminology requires that there be careful adjudication before the data is combined for analyses. Examples of data networking are the HMO Research Network and FDA's Sentinel Initiative.[52-54]

The advantage of these methods is the ability to create large datasets to study rare exposures and outcomes. Data pooling can be preferable to meta-analyses that combine the results of published studies because unified guidelines can be developed for inclusion criteria, exposures, and outcomes, and analyses using individual patient level data allow for adjustment for differences across datasets. Often, creation and maintenance of these datasets can be time consuming and expensive, and they generally require extensive administrative and scientific negotiation, but they can be a rich resource for CER.

Personal Health Records

Although they are not presently used for research to a significant extent, personal health records (PHRs) an alternative to electronic medical records. Typically, PHRs are electronically stored health records that are initiated by the patient. The patient enters data about his or her health care encounters, test results, and, potentially, responses to surveys or documentation of medication use. Many of these electronic formats are Web-based and therefore easily accessible by the patient when receiving health care in diverse settings. The application that is used by the patient may be one for which he or she has purchased access, or it may be sponsored by the health care setting or insurer with which the patient has contact. Other PHRs, such as HealthVault and NoMoreClipboard, can be accessed freely. One example of a widely used PHR is MyHealth*e*Vet, which is the personal health record provided by the VA to the veterans who use its health care system.[55] MyHealth*e*Vet is an integrated system in which the patient-entered data are combined with the EHR and with health management tools.

While there is ongoing research about how to best improve patient outcomes through the creative use of personal health records, there is also interest in how to best use the rich data contained within the personal health records for research. Outstanding issues remain regarding data ownership, but there is consensus that the data entered in the personal health record belongs to the patient and cannot be accessed without patient consent, which may include explicit documentation of the level of data-sharing that the patient would permit, at the time of entering data into the record. Many PHRs request that the patient state to whom he or she grants permission to access portions of the data.

Work is underway to standardize data collection across PHRs through the use of common terminologies such as the SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms) system. Presently, the National Library of Medicine (NLM) PHR project is validating and improving the NLM's clinical

vocabularies and studying consumers' use of PHR systems. In 2010, the NLM researchers reviewed and enhanced the controlled vocabulary for more than 2,000 condition names and synonyms and more than 300 surgery procedure names by enriching the synonymy, providing the consumer-friendly name when feasible, and adding SNOMED codes, when available, to these items.[56]

Patient-Reported Outcomes

Patient-reported outcomes (PROs) may occasionally be available in paper-based records and EHRs, but they are not presently found in administrative data. Wu et al. described several strategies that could be employed to increase the availability of PROs in administrative data.[57] The first is to encourage routine collection of PROs in clinical care by *requiring* it for compliance with data quality assurance guidelines. The Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey administered by CMS assesses patient's perspectives on their hospital care and could be a required activity. Another strategy, as described by Wu et al., is the required participation of all Medicare managed care plans with Medicare Advantage contracts in the Medicare Health Outcomes Survey, which collects data similar to that in the SF-12 Short-Form Health Survey. A third example may be provider reimbursement for collecting symptom-related outcome data, and thus its required reporting in administrative data. None of these approaches are currently widely used. Creative interventions to increase the availability of PROs in administrative data, ideally collected with validated tools and instruments, would be valuable to CER. Primary data collection of PRO information remains the most common means of ensuring that required PRO data are available on the patient population of interest at the required time points and of adequate completeness in order to conduct CER.

Conclusion

The choice of study data needs to be driven by the research question. Not all research questions can be answered with existing data, and some questions will thus require primary data collection. For questions amenable to the use of secondary data, observational research with existing data can be efficient and powerful. Investigators have a growing number of options from which to choose when looking for appropriate data, from clinical data to claims data to existing trial or cohort data. Each option has strengths and limitations, and the researcher is urged to make a careful match. In the end, the validity of the study is only as good as the quality of the data.

Checklist: Guidance and key considerations for data source selection for a CER protocol		
Guidance	**Key Considerations**	**Check**
Propose data source(s) that include data required to address the primary and secondary research questions.	– Ensure that the data resource is appropriate for addressing the study question. – Ensure that the key variables needed to conduct the study are available in the data source.	☐
Describe details of the data source(s) selected for the study.	– Nature of the data (claims, paper, or electronic medical records; if prospective, how the information is/was collected and from whom). – Coding system(s) that may be used (e.g., ICD9 or ICD10; HCPCS; etc.) – Population included in the data source (ages, geography, etc.). – Other features (e.g., health plan membership; retention rate [i.e., average duration of followup for members in the database, proportion of patients with followup sufficiently long for the study purpose]). – Time period covered by the data source(s). If non-U.S., describe relevant differences in health care and how this will affect the results.	☐
Describe validation or other quality assessments that have been conducted on the data source that are relevant to the data elements required for the study.	– If validation/quality assessments have not previously been performed, propose a method to assess data quality.	☐
Describe what patient identifiers are necessary for the research purpose, how they will be protected, and what permissions/waivers will be required.		☐
Provide details on any data linkage approach, and the quality/accuracy of the linkage, if applicable.	– Provide enough detail to clarify the quality of the linkage approach.	☐

HCPCS = Healthcare Common Procedure Coding System, ICD = International Classification of Disease

References

1. Bangsberg DR, Ragland K, Monk A, et al. A single tablet regimen is associated with higher adherence and viral suppression than multiple tablet regimens in HIV+ homeless and marginally housed people. AIDS. 2010 November 27;24(18):2835-40.
2. Brookhart MA, Schneeweiss S, Avorn J, et al. Comparative mortality risk of anemia management practices in incident hemodialysis patients. JAMA. 2010 March 3;303(9):857-64.
3. Zhang Y, Cotter DJ, Thamer M. The effect of dialysis chains on mortality among patients receiving hemodialysis. Health Serv Res. 2011 June;46(3):747-67.
4. Johansen KL, Zhang R, Huang Y, et al. Survival and hospitalization among patients using nocturnal and short daily compared to conventional hemodialysis: a USRDS study. Kidney Int. 2009 November;76(9):984-90.
5. Massarweh NN, Park JO, Yeung RS, et al. Comparative assessment of the safety and effectiveness of radiofrequency ablation among elderly Medicare beneficiaries with hepatocellular carcinoma. Ann Surg Oncol. 2012 Apr;19(4):1058-65.
6. Tan HJ, Wolf JS, Jr., Ye Z, et al. Complications and failure to rescue after laparoscopic versus open radical nephrectomy. J Urol. 2011 October;186(4):1254-60.
7. Canter D, Egleston B, Wong YN, et al. Use of radical cystectomy as initial therapy for the treatment of high-grade T1 urothelial carcinoma of the bladder: A SEER database analysis. Urol Oncol. 2011 September 7 [epub ahead of print].
8. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005 April;58(4):323-37.
9. Botsis T, Hartvigsen G, Chen F, et al. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Summits Transl Sci Proc. 2010 Mar 1;2010:1-5.
10. Hsiao C-J, Hing E, Socey TC. Electronic medical record/electronic health record systems of office-based physicians: United States, 2009 and preliminary 2010 state estimates. National Center for Health Statistics. 2009. Hyattsville, MD.
11. Federal Register. 74[11], 3328-332. 1-28-2009.
12. CPT® Process - How a Code Becomes a Code. www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt/cpt-process-faq/code-becomes-cpt.shtml. Accessed March 15, 2011.
13. Fisher ES, Whaley FS, Krushat WM, et al. The accuracy of Medicare's hospital claims data: progress has been made, but problems remain. Am J Public Health. 1992 February;82(2):243-8.
14. Quan H, Parsons GA, Ghali WA. Validity of procedure codes in International Classification of Diseases, 9th revision, clinical modification administrative data. Med Care. 2004 August;42(8):801-9.
15. Quan H, Parsons GA, Ghali WA. Assessing accuracy of diagnosis-type indicators for flagging complications in administrative data. J Clin Epidemiol. 2004 April;57(4):366-72.
16. Quan H, Li B, Saunders LD, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res. 2008 August;43(4):1424-41.
17. Segal JB, Ness PM, Powe NR. Validating billing data for RBC transfusions: a brief report. Transfusion. 2001 April;41(4):530-3.
18. Segal JB, Powe NR. Accuracy of identification of patients with immune thrombocytopenic purpura through administrative records: a data validation study. Am J Hematol. 2004 January;75(1):12-7.
19. Strom BL. Data validity issues in using claims data. Pharmacoepidemiol Drug Saf. 2001 August;10(5):389-92.
20. Thirumurthi S, Chowdhury R, Richardson P, et al. Validation of ICD-9-CM diagnostic codes for inflammatory bowel disease among veterans. Dig Dis Sci. 2010 September;55(9):2592-8.
21. Stein BD, Bautista A, Schumock GT, et al. The validity of ICD-9-CM diagnosis codes for identifying patients hospitalized for COPD exacerbations. Chest. 2012 Jan;141(1):87-93.
22. Tollefson MK, Gettman MT, Karnes RJ, et al. Administrative data sets are inaccurate for assessing functional outcomes after radical prostatectomy. J Urol. 2011 May;185(5):1686-90.
23. Wogen J, Kreilick CA, Livornese RC, et al. Patient adherence with amlodipine, lisinopril, or valsartan therapy in a usual-care setting. J Manag Care Pharm. 2003 September;9(5):424-9.

24. Yokoyama K, Yang W, Preblick R, et al. Effects of a step-therapy program for angiotensin receptor blockers on antihypertensive medication utilization patterns and cost of drug therapy. J Manag Care Pharm. 2007 April;13(3):235-44.
25. Hachem C, Morgan R, Johnson M, et al. Statins and the risk of colorectal carcinoma: a nested case-control study in veterans with diabetes. Am J Gastroenterol. 2009 May;104(5):1241-8.
26. Iqbal SU, Cunningham F, Lee A, et al. Divalproex sodium vs. valproic acid: drug utilization patterns, persistence rates and predictors of hospitalization among VA patients diagnosed with bipolar disorder. J Clin Pharm Ther. 2007 December;32(6):625-32.
27. JANUS, Data Standard Comparative Effectiveness Research. Available at: www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/ScienceBoardtotheFoodandDrugAdministration/UCM224277.pdf. Accessed October 23, 2012.
28. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. Ann Epidemiol. 1991 February;1(3):263-76.
29. Clark JM, Chang HY, Bolen SD, et al. Development of a claims-based risk score to identify obese individuals. Popul Health Manag. 2010 August;13(4):201-7.
30. National Cancer Institute SEER Training Module. Available at: http://training.seer.cancer.gov/registration/registry/history/. Accessed March 22, 2011.
31. Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. Annu Rev Public Health. 2011 April 21; 32:91-108.
32. Hammill BG, Hernandez AF, Peterson ED, et al. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. Am Heart J. 2009 June;157(6):995-1000.
33. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol. 1995 December 15;142(12):1255-64.
34. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. Am J Epidemiol. 2008 August 15;168(4):355-7.
35. Mackinnon A. The use and reporting of multiple imputation in medical research - a review. J Intern Med. 2010 December;268(6):586-93.
36. Foundation for the National Institues of Health. Observational Medical Outcomes Partnership. Available at: http://omop.fnih.org/node/22. Accessed May 8, 2011.
37. Gunn PP, Fremont AM, Bottrell M, et al. The Health Insurance Portability and Accountability Act Privacy Rule: a practical guide for researchers. Med Care. 2004 April;42(4):321-7.
38. Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. Annu Rev Med. 2006;57:575-90.
39. O'Keefe CM, Connolly CJ. Privacy and the use of health data for research. Med J Aust. 2010 November 1;193(9):537-41.
40. HIPAA Privacy Rule. National Institutes of Health Web site. http://privacyruleandresearch.nih.gov/. Accessed January 30, 2012.
41. Schneeweiss S, Setoguchi S, Brookhart A, et al. Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. CMAJ. 2007 February 27;176(5):627-32.
42. Boggon R, van Staa TP, Timmis A, et al. Clopidogrel discontinuation after acute coronary syndromes: frequency, predictors and associations with death and myocardial infarction—a hospital registry-primary care linked cohort (MINAP-GPRD). Eur Heart J. 2011 October;32(19):2376-86.
43. National Health Interview Survey. www.cdc.gov/nchs/nhis/about_nhis.htm. Accessed March 22, 2011.
44. Perlis TE, Des Jarlais DC, Friedman SR, et al. Audio-computerized self-interviewing versus face-to-face interviewing for research data collection at drug abuse treatment programs. Addiction. 2004 July;99(7):885-96.
45. Fairley CK, Sze JK, Vodstrcil LA, et al. Computer-assisted self interviewing in sexual health clinics. Sex Transm Dis. 2010 November;37(11):665-8.
46. Richens J, Copas A, Sadiq ST, et al. A randomised controlled trial of computer-assisted interviewing in sexual health clinics. Sex Transm Infect. 2010 August;86(4):310-4.
47. Corkrey R, Parkinson L. Interactive voice response: review of studies 1989-2000. Behav Res Methods Instrum Comput. 2002 August;34(3):342-53.

48. Abu-Hasaballah K, James A, Aseltine RH, Jr. Lessons and pitfalls of interactive voice response in medical research. Contemp Clin Trials. 2007 September;28(5):593-602.

49. Beasley JM, Kwan ML, Chen WY, et al. Meeting the physical activity guidelines and survival after breast cancer: findings from the after breast cancer pooling project. Breast Cancer Res Treat. 2012 January;131(2):637-43.

50. Herrinton LJ, Curtis JR, Chen L, et al. Study design for a comprehensive assessment of biologic safety using multiple healthcare data systems. Pharmacoepidemiol Drug Saf. 2011 Nov;20(11):1199-209.

51. Blettner M, Sauerbrei W, Schlehofer B, et al. Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int J Epidemiol. 1999 February;28(1):1-9.

52. Selby JV. Linking automated databases for research in managed care settings. Ann Intern Med. 1997 October 15;127(8 Pt 2):719-24.

53. Platt R, Davis R, Finkelstein J, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiol Drug Saf. 2001 August;10(5):373-7.

54. Harcourt SE, Smith GE, Elliot AJ, et al. Use of a large general practice syndromic surveillance system to monitor the progress of the influenza A(H1N1) pandemic 2009 in the UK. Epidemiol Infect. 2011 April 8;1-6.

55. Nazi KM, Hogan TP, Wagner TH, et al. Embracing a health services research perspective on personal health records: lessons learned from the VA My HealtheVet system. J Gen Intern Med. 2010 January;25 Suppl 1:62-7.

56. The Lister Hill National Center for Biomedical Communications. Annual Report FY 2010. Available at: www.lhncbc.nlm.nih.gov/lhc/docs/reports/2010/tr2010003.pdf. Accessed April 22, 2011.

57. Wu AW, Snyder C, Clancy CM, et al. Adding the patient perspective to comparative effectiveness research. Health Aff (Millwood) 2010 October;29(10):1863-71.