Good Research Practices for Comparative Effectiveness Research: Approaches to Mitigate Bias and Confounding in the Design of Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II

Emily Cox, PhD,[1] Bradley C. Martin, PharmD, PhD,[2] Tjeerd Van Staa, PhD, MD, MSc, MA,[3] Edeltraut Garbe, MD, PhD,[4] Uwe Siebert, MD, MPH, MSc, ScD,[5] Michael L. Johnson, PhD[6]

[1]Express Scripts, St. Louis, MO, USA; [2]Division of Pharmaceutical Evaluation and Policy, College of Pharmacy, University of Arkansas for Medical Sciences, Little Rock, AR, USA; [3]General Practice Research Database, London, UK; [4]Department of Clinical Epidemiology, Bremen Institute for Prevention Research and Social Medicine, Bremen, Germany; [5]Department of Public Health, Medical Decision Making and Health Technology Assessment, University of Health Sciences, Medical Informatics and Technology, Hall, Austria; Adjunct Professor of Public Health Policy and Management, Harvard University, Cambridge, MA, USA; [6]University of Houston, College of Pharmacy, Department of Clinical Sciences and Administration, Houston, TX, USA; Senior Scientist, Houston Center for Quality of Care and Utilization Studies, Department of Veteran Affairs, Michael E. DeBakey VA Medical Center, Houston, TX, USA

ABSTRACT

Objectives: The goal of comparative effectiveness analysis is to examine the relationship between two variables, treatment, or exposure and effectiveness or outcome. Unlike data obtained through randomized controlled trials, researchers face greater challenges with causal inference with observational studies. Recognizing these challenges, a task force was formed to develop a guidance document on methodological approaches to addresses these biases.
Methods: The task force was commissioned and a Chair was selected by the International Society for Pharmacoeconomics and Outcomes Research Board of Directors in October 2007. This report, the second of three reported in this issue of the *Journal*, discusses the inherent biases when using secondary data sources for comparative effectiveness analysis and provides methodological recommendations to help mitigate these biases.
Results: The task force report provides recommendations and tools for researchers to mitigate threats to validity from bias and confounding in measurement of exposure and outcome. Recommendations on design of study included: the need for data analysis plan with causal diagrams; detailed attention to classification bias in definition of exposure and clinical outcome; careful and appropriate use of restriction; extreme care to identify and control for confounding factors, including time-dependent confounding.
Conclusions: Design of nonrandomized studies of comparative effectiveness face several daunting issues, including measurement of exposure and outcome challenged by misclassification and confounding. Use of causal diagrams and restriction are two techniques that can improve the theoretical basis for analyzing treatment effects in study populations of more homogeneity, with reduced loss of generalizability.

Keywords: comparative effectiveness, epidemiology, nonrandomized studies, research design, secondary databases.

Background to the Task Force

In September 2007, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Health Science Policy Council recommended that the issue of establishing a task force to recommend Good Research Practices for Designing and Analyzing Retrospective Databases be considered by the ISPOR Board of Directors. The Council's recommendations concerning this new task force were to keep an overarching view toward the need to ensure internal validity and improve causal inference from observational studies, review prior work from past and ongoing ISPOR task forces and other initiatives to establish baseline standards from which to set an agenda for work. The ISPOR Board of Directors approved the creation of the task force in October 2007. Task force leadership and reviewer groups were finalized by December 2007 and the first teleconference took place in January 2008.

The task force members were experienced in medicine, epidemiology, biostatistics, public health, health economics, and pharmacy sciences, and were drawn from industry, academia and as advisors to governments. The members came from the UK, Germany, Austria, Canada, and the United States.

Beginning in January 2008, the task force conducted monthly teleconferences to develop core assumptions and an outline before preparing a draft report. A face-to-face meeting took place in October 2008, to develop the draft and three forums took place at the ISPOR meetings to develop consensus for the final draft reports. The draft reports were posted on the ISPOR website in May 2009 and the task forces' reviewer group and ISPOR general membership were invited to submit their comments for a 2-week reviewer period. In total, 38 responses were

Address correspondence to: Michael L. Johnson, University of Houston, College of Pharmacy, Department of Clinical Sciences and Administration, Houston, TX 77030, USA. E-mail: mikejohnson@uh.edu

10.1111/j.1524-4733.2009.00601.x

received. All comments received were posted to the ISPOR website and presented for discussion at the task force forum during the ISPOR 12th Annual International Meeting in May 2009. Comments and feedback from the forum and reviewer and membership responses were considered and acknowledged in the final reports. Once consensus was reached, the article was submitted to *Value in Health*.

Introduction

The goal of comparative effectiveness analysis is to examine the relationship between two variables, treatment or exposure, and effectiveness or outcome. The advantages of using secondary databases to examine this relationship are easily recognized by researchers in the field. Compared with data obtained through randomized controlled trials (RCTs), secondary data sources provide a low-cost means of answering the research question, answers can be obtained in a relatively short time frame, the data are more representative of routine clinical care and large cohorts of patients can be followed over long time periods [1]. However, researchers should be mindful of data limitations that, in some instances, preclude their use. In this section, we will address issues of validity with respect to secondary data sources and, where appropriate, provide researchers with tools to help mitigate threats to validity.

Researchers have been writing about the challenges that secondary data sources pose for more than two decades now [2–4], and although challenges still exist, the methodological approaches to address these challenges have greatly improved [5,6]. Key in contributing toward inaccuracies in administrative data is the fact that they were built for billing and record keeping purposes, not for research. Therefore, the potential for error occurs at many points along the record keeping process [7]. The implication for researchers is that both systematic and random error can occur in the identification of treatment exposure and outcome.

In RCTs, identifying and measuring exposure is done with a great deal of accuracy and precision. For example, in a clinical trial evaluation of drug treatment, not only is it known who has received the active drug, but also the degree of exposure—dose, duration and compliance with therapy. Similarly, outcomes—or measures of effectiveness—are measured with a great deal of accuracy and precision. Various devices and laboratory tests are used to measure and record both surrogate (blood pressure, cholesterol levels, tumor staging) and final end points (e.g., myocardial infarction [MI], stroke, and even death). This same level of precision is often not universally available in secondary data sources. Additionally, secondary data limit the measure of exposure and outcomes to those who seek care and is limited further in administrative claims data to those who obtain this care through the insurance payment system.

One way to measure the validity of exposure and outcomes using administrative data is to compare it with the gold standard. For outcome measures that gold standard is often patient self-report or the medical record. When using the gold standard of medical records, the sensitivity and specificity of medical claims data were found to have a high level of specificity but a great deal of variability in sensitivity across diagnoses [8]. For drug exposure, there have been indirect assessments of accuracy of prescription claims by comparing drug compliance metrics using pharmacy claims data with other compliance measures including patient self-report [9,10] and studies testing the accuracy of prescription claims information to define or supplement case definitions for hypertension [11]. Direct assessments of the validity of prescription claims comparing prescription claims with other medical data, such as a patients chart, have generally been performed in narrow populations or for selected drug classes and the results have been highly variable [4,12–15]. Although prescription claims are generally considered a valid measure of drug exposure, inaccuracies in measurement still exist.

Measurement of Exposure and Outcome

This paper first addresses how exposure and outcomes are measured using secondary data sources, discuss instances of misclassification and ways to mitigate these biases. This is followed by a discussion on confounding in epidemiological research and methodological approaches researchers should consider to control for confounding.

Measurement—Exposure

Secondary data sources measure drug exposure with varying degrees of accuracy. Table 1 highlights these data sources, the level of measurement, and inherent limitations in using these data sources for drug exposure.

Considered the most accurate and most commonly used measure of drug exposure is outpatient prescription claims. Prescription claims data provide a wealth of information on drug exposure including date of service, dispensing pharmacy, drug name, quantity, dose, and duration (days supply), and are considered by many to be the gold standard for measuring drug exposure [16]. It should be noted that days' supply can be unreliable for some drug classes (i.e., injectables or medications dosed on an as needed basis) and outside the United States, measures of duration may not be available.

Several options are available to identify drugs from outpatient prescription claims files. First is the National Drug Code (NDC), a 10-digit coding system established by the Food and Drug Administration (FDA) to uniquely identify drug, dosage, and package size. The FDA provides a complete listing of NDCs on its website however the drug lists can become cumbersome to manage, are time sensitive—changing with new drug entries or exits from the market. They can also be quite cumbersome to code particularly when a large number of NDCs codes are used. For example, using only the first nine digits of the NDC, which ignores package size, there are over 280 NDCs for the beta-

Table 1 Secondary data sources and measurement of drug exposure

Data source	Measurement	Type of exposure measured	Limitations
Outpatient prescription claims	NDC or therapeutic classification system (i.e., GCN, ATC, AHFS, etc.)	Incidence and prevalence use and intensity of exposure	See expanded discussion on misclassification
Medical records/charts	Drug name, dosage and regimen for prescribed and OTC agents	Binary drug exposure (incidence/prevalence)	Incomplete capture of patients' medication history; does not capture degree of exposure (i.e., duration)
Outpatient medical claims	Health Care Procedure Codes only for select medications	Binary drug exposure (incidence/prevalence) and persistency	Limited to only those medications administered in the physician's office

AHFS, American Hospital Formulary Service; ATC, Anatomical Therapeutic Chemical; GCN, generic code number.

Table 2 Secondary data sources for measuring outcomes

Data source	Measurement	Outcome measured
Medical records	Manual or automated (electronic medical records) extrapolation of diagnoses, procedures and treatments, biomarkers and other laboratory data	Used alone or with other data sources to identify disease progression, surrogate, or final end points
Outpatient medical claims	ICD-9 or ICD-10-CM, OXMIS, This paper first addresses how exposure and outcomes are measured using secondary data sources, discuss instances of misclassification and ways to mitigate these biases. This is followed by a discussion on confounding in epidemiological research and methodological approaches researchers should consider to control for confounding, CPT-4, OPCS-4, laboratory testing, diagnostic tests	Used alone or with other data sources to identify disease progression, surrogate or final end points
Eligibility files		Death*
Inpatient medical claims	ICD-9 or ICD-10-CM, OXMIS, The Read Codes, CPT-4, OPCS-4, laboratory testing, diagnostic tests, discharge status	Used alone or with other data sources to identify disease progression, events or final end points

*May not be documented as such in all cases.
ICD, International Classification of Diseases; OXMIS, Oxford Medical Information System; CPT, current procedural terminology; OPCS-4, Office of Population, Censuses and Surveys Classification of Operations and Surgical Procedures (4th revision).

blocker Atenolol. To simplify drug identification, researchers can purchase a therapeutic classification system such as the American Hospital Formulary Service Pharmacologic-Therapeutic Classification, Red Book, Anatomical Therapeutic Chemical classification system or Medi-Span's Generic Product Identifier. These systems link NDC to drug classes, which allow for more manageable coding of unique drugs or therapy class.

Medical records can be another data source to identify drug exposure recording whether the physician prescribed medication for the patient, the dose, and intended regimen. However, medical records do not record whether the patient obtained the medication from the pharmacy, or typically the degree of exposure (i.e., compliance). Additionally, the medical record (either inpatient or outpatient) does not record all prescribed medications taken by patients and is generally not considered a valid source for identifying drug exposure. However, medical records may be considered as a source for capturing over-the-counter (OTC) agents, typically not covered or captured in the prescription claim record. In the United States, it should be noted that the FDA will not accept e-medical records as a source for measuring drug exposure.

Drug exposure can also be measured using outpatient medical claims for a limited number of medications dispensed and administered in the physicians' office. In the United States, these are captured using Health Care Procedure Codes (HCPCS). However, drug use identified from HCPCS codes do not indicate dosage and are not immediately assigned to newer agents. Additionally, medical billers often use miscellaneous J-codes when billing for medications administered in physician offices, which does not allow for accurate identification of the drug administered.

Another challenge faced by researchers in measuring exposure is accounting for switching in the assignment to exposure groups. Switching from one drug therapy to another often occurs naturally as a result of treatment failure or systematically from changes in benefit design [17] or programmatic features such as formulary status changes. Researchers should establish criteria a priori for treatment group assignment, be transparent in methods, and conduct sensitivity analysis to determine the impact of treatment identification on study results.

Measurement—Outcomes

For a given disease or condition, various measures of clinical effectiveness exist. For example, in the treatment of high cholesterol, measures of clinical effectiveness include both intermediate measures, such as the biomarker low-density lipoprotein cholesterol and cardiovascular end points including stroke or MI. Outside of the RCT environment, researchers face limitations in measuring effectiveness, particularly those that involve intermediate biomarkers or self-reported symptom scales and measures of patient functioning. Among secondary data sources, medical records are typically considered the gold standard for capturing intermediate and final outcomes (Table 2). Other secondary data sources, although providing a wealth of information on treatment patterns and medical events are more limited in measuring effectiveness. Administrative claims data can identify final end points such as fractures, stroke, or MI but are limited to proxy measures at best in the measurement of intermediary outcomes. Using a combination of diagnostic, procedure, or facility codes, researchers are beginning to develop proxy measures of intermediate outcomes with some success. For example, a study examining disease severity for chronic obstructive pulmonary disease used diagnostic and inpatient hospital stays to classify severe or moderate COPD and found moderate accuracy to medical charts [18].

There is growing use of laboratory results data linked to administrative claims data to measure intermediate outcomes. However, these data are as yet to be made available on a large scale in the United States.

Classification Bias

Systematic and random errors can occur in measuring both exposure and outcome resulting in the violation of internal validity. This error is termed classification bias—identifying subjects as being exposed to drug when they are not or not exposed when they are. Classification bias is further categorized as differential or nondifferential and unidirectional or bidirectional. Nondifferential misclassification occur when the likelihood of misclassification is the same across the exposed or outcome groups. For example, exposure misclassification for a low-cost medication using prescription claims data would be equally likely regardless of outcome. However, differential misclassification is present when the likelihood of misclassification is different between exposed or outcome groups. An example of differential misclassification for drug exposure is when those who are exposed have a lower likelihood of outcome misclassification because to receive medication they have to enter the health-care system, which increases their likelihood of recording a diagnosis. Those not exposed are much more likely to be misclassified as not having the disease, which is an artifact of not entering the health care system. Unidirectional misclassification occurs when the direction of the misclassification is in the same direction. Bidirectional misclassification occurs when the likelihood of misclassification is in both directions—there is a probability that cases appear as controls and controls appear as cases. For a more complete discussion see Hartzema and Perfetto [19]. As a researcher, one should consider and state the direction of potential sources of misclassification and how that could influence the rejection of the null hypothesis [7].

An important data element influencing classification bias of both drug exposure and outcomes when using secondary data sources is member eligibility. In the United States, many administrative datasets are linked to employment and natural transitions in the labor market can influence classification bias. If eligibility is not accounted for in the measure of medication compliance, for example, those not continuously eligible may be incorrectly classified as noncompliant when in fact the lack of drug exposure was caused by the loss of eligibility. Statistically controlling for length of eligibility or limiting to continuously eligible in these instances may be most appropriate. Lack of appropriate time for follow-up because of drops in eligibility is also a concern for outcomes misclassification if member follow-up does not allow for capture of the clinical event. Eligibility must be controlled for and lack of this information precludes comparative effectiveness research.

Drug Exposure Misclassification

Many factors can lead to misclassification with respect to drug exposure. With outpatient prescription claims, a greater number of opportunities for misclassification in the direction of not exposed exist given the multiple channels by which members can receive their medications outside of the reimbursement arrangements of third-party payers. Other means for obtaining prescription drugs that would preclude claims capture include physician samples, patient assistance programs (PAP), paying out of pocket, inpatient hospital stays, taking a medication belonging to someone else, secondary insurance coverage, or fraudulent behavior. The likelihood of this misclassification can be influenced by patient demographics and plan design. For example, the elderly and lower-income patients or those facing higher out-of-pocket payments may be more likely to participate in PAP programs or obtain samples from their physician, leading to systematic misclassification.

Various trend and utilization management programs can also lead to misclassification. Programmatic features including prior authorization policies, caps, or maximum limits on coverage, and pharmaceutical step therapy programs can influence not only the measure of exposure but assignment to exposure category. This information, although not always readily available to researchers, would represent a major limitation if not documented. More recently, the proliferation in the United States of no-cost or low-cost generic programs offered by retail chain pharmacies are increasing the likelihood of misclassification since these claims are not captured by the health plan. This could lead to bias depending upon the drug comparators, study sample, or geographic region given that market penetration of these programs differ by region.

Differences in formularies, or the list of covered drugs, can lead to misclassification. Systematic errors in exposure classification can occur when the treatments being compared have different formulary status or are on different tiers. If drug A is a second tier product being compared with drug B, which is a third tier product where members pay a higher copayment, differential classification bias could result, assuming higher copayments lead to lower compliance, which could impact outcomes. Methods to address these issues are covered in later sections.

Additionally, for administrative claims data, classification bias is present when measuring exposure for OTC medications or medications with limits or coverage exclusions (medications used to treat cosmetic indications). For ambulatory comparative effectiveness analysis, hospital stays (or other inpatient stays) must be accounted for in the statistical analysis [20]. However, the random bias that occurs when patients use other channels to receive medication can only be addressed as a potential study limitation.

The level of exposure misclassification can also be influenced by the study design. One important choice in the design of database studies is the time-window during which patients are considered "exposed." This will impact misclassification of not only exposure but also outcome measurement. A study that is based on prescription information can use, for example, a 3-month time-period following each prescription in order to assess the outcome and estimate the risk of the outcome during this time-period. Although this is not always recognized, the choice of this exposure time-window is of major importance. Since misclassification of the relevant exposure time will lead to a nondifferential bias toward the null. The choice of the exposure time-window should not be based on the actual drug intake, but rather on the time-period during which the medication may cause the outcome and the duration of the pathogenic process [21,22].

As an example, a study of the effects of a medication on the risk of malignancies may suffer from a major exposure misclassification if the exposure time window would be based on the time period of drug intake and the study would include many short-term users. On the other hand, a study of allergic reactions would also suffer from exposure misclassification if the exposure time window goes beyond what is considered clinically relevant. Different approaches to improve the characterization of the exposure time window include efforts to validate the relationship or sensitivity analysis, repeating the analysis with different exposure time windows.

The focus of exposure misclassification has been in the direction of not exposed. However, it should be noted that the direction of misclassification can also be toward exposure. It cannot be assumed that presence of a claim indicates that the patient actually took the medication. For example, patients may obtain a medication for antibiotic or pain therapy, and take only if symptoms appear increasing the likelihood for misclassification toward exposed.

Outcome Misclassification

Several factors can lead to misclassification of diagnostic or procedure codes including plan payment systems, diagnoses, and the specificity of coding in the database [1,7]. Reimbursement systems based upon capitated payment arrangements where providers are less incentivized to submit claims documenting care compared with fee for service payment arrangements are more prone to classification bias. Under capitated payment systems, researchers should proceed with caution and attempt to validate claims data with external data sources (i.e., medical chart review).

Misclassification has been shown to vary by disease state with hypertension and diabetes having the highest rates of sensitivity (60.6 and 62.6, respectively) and chronic liver disease, peptic ulcer disease, or acute myocardial infarction with some of the lowest levels of sensitivity (27.6, 27.6, and 25.4, respectively) [8]. This variability can be caused by multiple factors including clinical ambiguity in diagnoses, stigma associated with the diagnoses or coding used for rule out diagnostic procedures. Using a longer look back period and requirements of at least two diagnoses or inclusion of medical treatment can increase specificity [23]. Also being explored is the use of algorithms using drug, medical and patient demographic information to increase the accuracy of diagnostic information [24]. Systematic error in classification of outcomes can occur if the researcher fails to take into account changes in codes resulting from updates or brought about by the transition from the International Classification of Diseases (ICD)-9 to ICD-10 coding systems.

When considering various approaches, researchers should seek out definitions that have been validated with external sources, such as chart review. When there are several approaches without a clear empirical direction, sensitivity analyses should be explored to understand the implications of the various definitions on the results. For example, MI may be defined using two diagnoses or one diagnosis and a hospital stay, which will alter the incidence of MI detected in the study. When measuring comorbidity ideally one should select a measure that has been validated in a population most similar to the study and for the outcome under investigation.

Recommendations

1. State the direction of potential sources of misclassification and how that could influence the acceptance or rejection of the null hypothesis.
2. Eligibility must be controlled for and lack of this information precludes comparative effectiveness research.
3. For ambulatory comparative effectiveness analysis, hospital stays (or other inpatient stays) must be accounted for in the statistical analysis.
4. The choice of the exposure time-window should not be based on the actual drug intake, but rather on the time-period during which the medication may cause the outcome and the duration of the pathogenic process.
5. Definitions that have been validated with external sources, such as chart review, should be used as the primary method in defining the measure. When there are several approaches without a clear empirical direction, sensitivity analyses should be explored to understand the implications of the various definitions on the results.
6. When measuring comorbidity, select a measure that has been validated in a population most similar to the study and for the outcome under investigation.

Confounding and Causal Graphs

Issues surrounding misclassification is not the only bias that researchers are faced with when using retrospective secondary data sources. Confounding also comes into play. Confounding is classically defined as a bias that distorts the exposure-disease or exposure-outcome relationship [25]. Frequently used definitions of confounding and standard textbook methods to control for confounding state that a confounder is an independent (causal) risk factor for the outcome of interest that is associated with the exposure of interest in the population, but that is not an intermediate step in the causal pathway between the exposure and the outcome [26,27].

Confounding by Indication for Treatment

A common and pernicious problem endemic to pharmacoepidemiologic studies is confounding by indication of treatment. For example, when the choice of therapy is affected by the severity of illness, and physicians prescribe one therapy over another depending on the severity and the perceived effectiveness of one drug compared with another for patients with differing severity levels, then confounding by indication for treatment occurs (assuming that the severity of disease also is a risk factor for the outcome of interest). In this case, apparent (i.e., estimated) treatment effects are confounded, that is, they are not causal but they may actually be caused by the severity of illness that led to patients being prescribed a given treatment.

Measured versus Unmeasured Confounding

Confounders may be measured or unmeasured. Secondary databases of a variety of sources may contain a wide and rich variety of information that can be used to measure an array of potentially confounding factors. However, even the most detailed and complete data sources may fail to include information on potential confounding factors, and these remain unmeasured and hence uncontrolled in a given study leading to residual confounding. Methods to address both measured and unmeasured (residual) confounding factors have been developed to address these concerns and will be detailed in the third series of the task force's report.

Time-Dependent Confounding

The more complicated (but probably not less common) case of *time-dependent* confounding refers to variables that simultaneously act as confounders and intermediate steps, that is, confounders and risk factors of interest mutually affect each other. Confounding by indication, may take the form of time dependent confounding. An example is the effect of aspirin use (treatment) on risk of MI and cardiac death (outcome). Prior MI is a confounder for the effect of aspirin use on risk of cardiac death, because prior MI is a cause of (subsequent) aspirin use, and is also a causal risk factor for (subsequent) cardiac death. However, (prior) aspirin use also causally prevents prior MI. Therefore, prior MI simultaneously acts as confounder (causing aspirin use) and intermediate step (being affected by aspirin use), and hence is a time-dependent confounder affected by previous treatment.

Traditional textbook techniques to control for time-independent confounding include restriction, stratification, matching, or multivariate regression analysis. However, these methods have been criticized for being inadequate to control for time-dependent confounding. Other methods such as g-computation, marginal structural models, or structural nested models have been suggested as approaches to this problem [28,29].

These analytic methods require repeated measurements of the treatment of interest, potential confounders and the outcome. With the proliferation of longitudinal data sources, where patients are followed up over years of exposure to medical therapies, these analytic methods should be applied.

Causal Graphs

To address the issue of confounding in retrospective databases and to be able to do a proper causal analysis, we must answer three questions: 1) which a priori assumptions can we make about the causal relationships between the variables of an epidemiological study?; 2) under these assumptions, are the observed data sufficient to control for confounding?; and 3) what methods are appropriate to control for confounding?

Causal graphs can guide us in answering these questions [30]. Directed acyclic graphs (DAGs) are causal graphs that can be used to understand and explicitly state causal a priori assumptions about the underlying biological mechanisms [31,32]. DAGs consist of a set of nodes and directed links (arrows) that connect

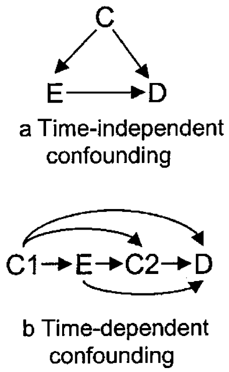

Figure 1 Simple directed acyclic graph showing (a) time-independent and (b) time-dependent confounding.

certain pairs of nodes (see Fig. 1). For our purposes, nodes represent variables and arrows denote causal relationships. A set of precise graphical rules for DAGs has been developed, which allows us to determine whether an unbiased effect is estimable from the observed data, which variables must be adjusted for in the analysis, and which statistical methods can be used to obtain unbiased causal effects. Part of these rules is a new and graphically oriented definition of confounding (i.e., the "backdoor criterion").

Furthermore, DAGs offer a readily accessible approach to understanding complex statistical issues including the fallibility of estimating direct effects (i.e., controlling for intermediate steps), the rationale for instrumental variables, and controlling for compliance in randomized clinical trials (when both "intention-to-treat" [ITT] and "per protocol" analyses can fail to yield the true causal intervention effect). In conclusion, DAGs are a valuable and comprehensive tool that offers epidemiologists and outcomes researchers better insight into confounding and the causal interpretation of their model results.

Another example of time-dependent confounding by treatment is antiviral treatment of HIV infection, where treatment or dose may depend on CD4-count and this dependency may continue over the course of the disease [33].

In the conduct of nonrandomized comparative effectiveness studies, it is strongly recommended to: 1) define the DAG for the base-case analysis before actually starting the analysis; 2) report the DAG for the base-case analysis; and 3) if sensitivity analyses are performed for different assumptions regarding the confounding structure, to report the additional DAGs representing the assumptions of the respective sensitivity analyses.

One additional approach for assessing the likelihood that residual confounding may be responsible for an observed treatment effect would be to conduct a sensitivity analysis exploring the effect of the treatments on outcomes that should not be influenced by the treatment in addition to the primary end points. For example, in an analysis to compare different prostate cancer treatments, one could explore recurrent cancer related mortality or all cause mortality as primary end points and additionally compare the treatments on outcomes unrelated to the disease or the treatments such as pulmonary or diabetes-related mortality. If a prostate cancer treatment was found to have a beneficial effect on cancer mortality but no impact on diabetes related mortality, there is greater confidence linking the treatment to the primary end point, however, if a beneficial effect is also found for diabetes mortality, it is less clear if the treatments are influencing the primary cancer benefit or if the therapy is more likely to be prescribed for "healthier" patients that could not be controlled in the analysis [34].

Recommendations:

1. Define the DAG for the base-case analysis before actually starting the analysis.
2. Report the DAG for the base-case analysis.
3. If sensitivity analyses are performed for different assumptions regarding the confounding structure, report the additional DAGs representing the assumptions of the respective sensitivity analyses

Restriction—Inclusion and Exclusion Criteria

Although a variety of systematic errors may bias nonexperimental research [35] confounding bias is of particular concern in epidemiologic studies of drug effects [36].

Restricting study cohorts to patients who are homogeneous regarding their indication for the study drug will lead to more balance of patient predictors of the study outcome among exposure groups and thus will reduce confounding but not necessarily eliminate confounding, particularly when there are variables that influence prescribing decisions that are not available in the data. Restricting study cohorts can also increase the likelihood that all included subjects will have a similar response to therapy and therefore reduce the likelihood of effect modification. RCTs commonly restrict their study population to patients with a presumed indication for the study drug and then randomly allocate the actual treatment.

There are many different approaches to restriction in specific studies [37] and it is therefore difficult to provide generic advice that fits specific study designs. However, several guiding principles can be identified that should be considered in a nonrandomized database study on effectiveness and safety of medical interventions [38].

Exclude Patients with a History of the Study Outcome?

The decision whether to exclude patients with a history of the study outcome is largely based on the study questions and the chronicity of the outcome under study. Some guiding principles may include:

1. Patients with a history of occasionally or frequently occurring events that are restored to a normal health level with or without treatment may not be candidates for exclusion if their health status has reached a normal level before cohort entry. Examples for such conditions are uncomplicated viral or bacterial infections.
2. Patients with a history of conditions that are markers for an underlying chronic condition will have an increased risk for the study outcome and at the same time may be more likely to take a study medication causing confounding. Examples for such conditions include hip fractures in elderly patients, which are markers for frail health and/or osteoporosis, which put the patient at increased risk for a future event. Similarly, a previous MI is a strong risk factor for future cardiac events. If these conditions are strong risk factors for future events and therefore potentially strong confounders it may be better to exclude these patients from the analysis rather than adjusting for them.

Study Incident Medication Users Only?

Usually, an epidemiologic database study is implemented by defining a study period for which subjects are considered. Let us consider a cohort study of statin use and some health outcome. The most basic cohort definition would be to identify subjects who used a statin at any point during the study period, assigning the date of first observed statin use during that period as an index date. On each statin user's index date, we sampled a subject who had not used a statin as of that date, i.e., a nonuser, and assigned him or her the same index date.

The population of statin users described earlier consists of a mix of incident drug users, i.e., those starting on a statin, and prevalent users, i.e., those taking a statin for some time.

Mixed Prevalent and Incident User Cohorts

Studying mixed prevalent and incident user cohorts will lead to under-ascertainment of early events. Depending on the average duration (chronicity) of use, such cohorts may be composed predominantly of prevalent users and few new users (e.g., statins). The estimated average treatment effect will therefore underemphasize effects related to drug initiation and will overemphasize effects of long-term use [39].

Prevalent users of a drug have by definition persisted in their drug use, similar to the concept of survivor cohorts in chronic disease epidemiology [40]. Being persistent or adherent is a characteristic found more frequently in patients who tolerate the drug well and who perceive some therapeutic benefit. Adherence also characterizes patients with higher educational status and health-seeking behavior particularly if the study drug is treating an asymptomatic condition like statins treating hyperlipidemia, characteristics that are difficult to assess in claims data, and may lead to healthy user bias [41–43].

The duration of use among prevalent users can differ by drug exposure; duration thus may cause bias if it remains unadjusted. Such a scenario is likely when newly marketed drugs are compared with competitors that have been available longer. In database studies, duration of prior use can only be assessed by tracing back a continuous string of prescriptions to the initial prescription.

In studying prevalent users, investigators can assess patient characteristics only after the initial exposure; thus the drug under study may affect those characteristics. Adjusting for such factors that are on the causal pathway of the drug's action will lead to an underestimation of the drug effects.

"New User Design." One begins an incident user design by identifying all patients in a defined population who start a course of treatment with the study medication. Exposed person-time begins at the start of treatment, which is identified as a dispensing of the index drug without a dispensing of the index drug during the prior year or some other fixed time interval comparable with a wash-out period commonly used in RCTs. The advantage of the so-called "New User Design" has recently been summarized [40]. Although limiting the study population to drug initiators resembles one of several key characteristics of clinical trials, the limited number of incident users requires large source populations like health care utilization databases from which new starters can be identified efficiently. For some patients it may not be the first time they take the study drug, i.e., they are not really naïve to the drug. Patients who know from earlier treatment courses that they tolerate the drug and that it is effective for them are more likely to use the same drug again. The chance of an initiator to be a true new user can be increased by requiring longer periods without use of the study drug before the index prescription.

What Is the Most Adequate Comparison Group?

Choosing a comparison group is a complex and sometimes subjective issue. The ideal comparison should comprise patients with identical distributions of measured and unmeasured risk factors of the study outcome.

Patients with the same treatment indication: "Alternative Drug Users." Selecting comparison drugs that have the same perceived medical indication for head-to-head comparisons of active drugs will reduce confounding by selecting patients with the same indication (e.g., indication for using celecoxib vs. rofecoxib). Although one can rarely measure the indication directly—in the statin example we would need laboratory values of serum lipid levels that are not routinely available in claims data—we infer the indication by the initiation of a treatment specific to the indication. When studying unintended benefits or risks of drugs, such as exploring the potential cancer preventive properties of nonsteroidal anti-inflammatory drugs (NSAIDs), confounding by indication may be less problematic as physicians are unlikely to prescribe therapies based on a patient's risk of developing the un-intended outcome (e.g., cancer) assuming the disease(s) for which the therapy are indicated are unrelated to the outcome. However, new competitors within a class are often marketed for better efficacy, slightly expanded indications, or better safety (cyclo-oxygenase-2 inhibitors [coxibs] vs. nonselective NSAIDs) influencing physicians' prescribing decisions [44]. In this way, new opportunities of confounding by indication can arise.

"Nonusers." In some cases there either is no comparator drug with a reasonably close indication to the study drug or a class effect is suspected such that the entire class is to be tested, requiring comparison subjects who did not use any drug of this class. The most obvious choice may be to identify study subjects who do not use the study drug and then to pick a random date as the index date, possibly matched by time to the index date of the first prescription among active drug users.

Obviously, patients on therapy most likely have a medical indication; by contrast a large proportion of nonusers have no medical indication, i.e., patients initiating statin therapy are more likely to have elevated lipid levels and therefore increased cardiac risks. However, nonusers as defined earlier may differ substantially from users of the index drug for both measured and unmeasured characteristics, even beyond the indication for the index drug.

As a case in point: Although initiators of a new drug have (presumably) been evaluated by a physician just before that prescription, nonusers may not have seen a physician for a while and, in fact, may have less contact with the health care system in general. Differential underrecording of health conditions in the nonuser comparison group makes members of the comparison group appear healthier than they really are and may lead to an overestimation of treatment effects.

Groups will be more comparable regarding access to health care, including health-seeking behavior and disease surveillance, when choosing comparison patients who also had contact with the health-care system in the form of a drug dispensing. Like patients starting the study drug, such patients have just been evaluated by a physician before the initial prescription. Adequate comparison groups for new statin initiators could, for example, be initiators of topical glaucoma drugs or thyroid hormone substitution. Both these classes of pharmaceuticals are unrelated to lowering serum lipid levels and are used for preventing the progression of an initially asymptomatic condition.

Excluding Patients with Contraindications?

In studies of the effectiveness of drugs it is questionable whether we want to include patients who have a clear contraindication to the study drug. Such patients will be few and their experience will be unusual. Prudence dictates, therefore, excluding patients with contraindications or absolute indications, resulting in a situation similar to the therapeutic equipoise required for RCTs [45].

Because reliably identifying contraindications in claims data is unlikely, identifying them empirically is more promising. Propensity scores, a common mechanism for doing this, estimate each patient's probability of treatment given all measured covariates. These propensity scores follow a distribution between 0 and 1 that differ between actual users and nonusers. On the low end of the propensity score distributions indicating a low propensity for receiving treatment, there will be a range that is only populated by actual nonusers because all users have a higher propensity scores. Such nonusers are likely to have a contraindication for the study medication because no subject with such a low propensity score has actually received treatment. These patients should be deleted from the study population. Analogously, such trimming can be considered at the upper end of the propensity score, excluding patients who will always be treated.

Excluding Patients with Very Low Adherence?

Patients dropping out of RCTs for reasons related to the study drug may cause bias. Noninformative dropout causes bias towards the null in ITT analyses. The medical profession and regulatory agencies accept such a bias because its direction is known and trial results are considered conservative regarding the drug's effectiveness. Discontinuation of treatment may also be associated with study outcomes. Obvious reasons are lack of perceived treatment effect or intolerance. Both factors may lead to early stopping but can cause discontinuation at any time later during the course of treatment. Another factor that may lead to discontinuation of medications, particularly those used to treat asymptomatic conditions, is overall frail health status that requires multiple medications to treat the more symptomatic conditions. For example, cancer patients may discontinue statins in order to reduce polypharmacy in favor of more urgently needed drugs [42].

RCTs try to minimize bias from nonadherence by frequently reminding patients and by run-in phases before randomization aimed to identify and exclude nonadherent patients. In routine care, adherence to drugs is unfortunately substantially lower than in RCTs. Studies have shown, that for statin medications, only 50% to 60% of elderly patients refill their prescriptions after 6 months [46].

Starting follow-up after the third fill of a chronic medication will exclude patients who are least adherent. Unlike RCTs in which run-in phases are often done with placebo [47] patients in routine care experience their first exposure to a new drug and may discontinue use because of a lack of effectiveness or intolerance during what may be the most vulnerable period for some medication-outcome relations. As long as that proportion is small and most patients discontinue for reasons not directly related to the study drug(s), this issue should be minor.

Generalizability

To guide our thinking about generalizability, it is useful to specify the patient to whom we wish to generalize our results. From a patient and physician perspective, the most relevant and frequently asked question is, "What is the effectiveness and safety of a particular drug that I am about to start and continue to use, compared with not starting therapy, or compared with starting an alternative drug?" From this viewpoint, restricting studies to initiators of drug therapy does not limit generalizability. Instead, it avoids under-representation of treatment effects that occur shortly after initiation. Patients with known contraindications (or their clinicians) would usually not have to confront this hypothetical question because prescribing the drug in the first place would contravene current medical knowledge. Therefore, excluding patients with known contraindications places little limits on generalizability.

In making a prescribing decision, physicians must assume that patients will take a drug as directed. If clinicians knew beforehand that a patient would not take a prescribed medication, they would not ponder the appropriateness of the drug in the first place. Consequently, excluding patients who are nonadherent to their treatment independent of intolerance or treatment failure—will not limit generalizability to the question raised above. However, the situation is quite different if we restrict the study population by disease severity, comorbidities, polypharmacy, and other risk factors for the study outcome. Data based on such restrictions will limit physicians when making prescribing decisions concerning the excluded patient subgroups. The obvious solution to this problem is to stratify analyses according to relevant clinical subgroups, rather than restricting them out of the analysis altogether, and then testing whether treatment effects differ between groups [48]. The large size of health-care utilization databases can allow performing such subgroup analyses with substantial numbers of subjects, and represents an attractive alternative to wholesale restriction.

Conclusion

Design of nonrandomized studies of comparative effectiveness face several daunting issues, including measurement of exposure and outcome challenged by biases in misclassification and confounding. We identified a set of restrictions that analysts should consider in studies of the effectiveness of therapies when using large observational databases. Such restrictions will place few limits on generalizability of research finding for most clinically relevant treatment choices. Use of causal diagrams and restriction are two techniques that can improve the theoretical basis for analyzing treatment effects in study populations of more homogeneity, with reduced loss of generalizability.

Source of financial support: This work was supported in part by the Department of Veteran Affairs Health Services Research and Development grant HFP020-90. The views expressed are those of the authors and do not necessarily reflect the views of the Department of Veteran Affairs.

References

1 Schneeweiss S. Developments in post-marketing comparative effectiveness research. Clin Pharmacol Ther 2007;82:143–56.

2 Roos LL, Sharp SM, Wajda A. Assessing data quality: a computerized approach. Soc Sci Med 1989;28:175–82.

3 Motheral BR, Fairman KA. The use of claims databases for outcomes research: rationale, challenges, and strategies. Clin Ther 1997;19:346–66.

4 Tamblyn R, LaVoie G, Petrella L, Monette J. The use of prescription claims databases in pharmacoepidemiological research: the accuracy and comprehensiveness of the prescription claims database in Quebec. J Clin Epidemiol 1995;48:999–1009.

5 Berger M, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting non-randomized studies of treatment effects using secondary data sources. ISPOR TF Report 2009—Part I.

6 Johnson ML, Crown W, Martin B, et al. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from non-randomized studies of treatment effects using secondary data sources. ISPOR TF Report 2009—Part III.

7 Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323–37.

8 Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol 2004; 57:131–41.

9 Erickson SR, Coombs JH, Kirking DM, et al. Compliance from self-reported versus pharmacy claims data with metered-dose inhalers. Ann Pharmacother 2001;35:997–1003.

10 Grymonpre RE, Didur CD, Montgomery PR, et al. Pill count, self-report, and pharmacy claims to measure medication adherence in the elderly. Ann Pharmacother 1998;32:749–54.

11 Quam L, Ellis LB, Venus P, et al. Using claims data for epidemiologic research. The concordance of claims-based criteria with the medical record and patient survey for identifying a hypertensive population. Med Care 1993;31:498–507.

12 Kwon A, Bungay KM, Pei Y, et al. Antidepressant Use: concordance between self-report and claims records. Med Care 2003;41:368–74.

13 McKenzie DA, Semradek J, McFarland BH, et al. The validity of Medicaid pharmacy claims for estimating drug use among elderly nursing home residents: the Oregon experience. J Clin Epidemiol 2000;53:1248–57.

14 Kirking DM, Ammann MA, Harrington CA. Comparison of medical records and prescription claims files in documenting prescription medication therapy. J Pharmacoepidemiol 1996;5:3–15.

15 King MA, Purdie DM, Roberts MS. Matching prescription claims with medication data for nursing home residents: implications for prescriber feedback, drug utilization studies and selection of prescription claims database. J Clin Epidemiol 2001;54:202–9.

16 Strom BL, Carson JL, Halpern AC, et al. Using a claims database to investigate drug-induced Stevens-Johnson syndrome. Stat Med 1991;10:565–76.

17 Mager DE, Cox ER. Relationship between generic and preferred-brand prescription copayment differentials and generic fill rate. Am J Manag Care 2007;13:347–52.

18 McKnight J, Scott A, Menzies D, et al. A cohort study showed that health insurance databases were accurate to distinguish chronic obstructive pulmonary disease from asthma and classify disease severity. J Clin Epidemiol 2005;58:206–8.

19 Hartzema AG, Perfetto EM. Sources and effects of drug exposure and unintended effect misclassification in pharmacoepidemiologic studies. In: Hartzema AG, Porta MS, Tilson HH, eds. Pharmacoepidemiology (2nd ed.). Cincinnati, OH: Harvey Whitney Books Co., 1991.

20 Suissa S. Immeasurable time bias in observational studies of drug effects on mortality. Am J Epidemiol 2008;168:329–35.

21 Van Staa TP, Abenhaim L, Leufkens HGM. A study of the effects of exposure misclassification due to the time-window design in pharmacoepidemiologic studies. J Clin Epidemiol 1994;47: 183–9.

22 Van Staa TP, Abenhaim L. Utilization dynamic and risk comparisons in studies that use prescription information. Pharmacoepidemiol Drug Saf 1994;3:191–7.

23 Kiyota Y, Schneeweiss S, Glynn RJ, et al. The accuracy of Medicare claims-based diagnosis of acute myocardial infarction: estimating positive predictive value based on review of hospital records. Am Heart J 2004;148:99–104.

24 Lix LM, Yogendran MS, Leslie WD, et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. J Clin Epidemiol 2008;61:1250–60.

25 Miettinen OS. Confounding and effect modification. Am J Epidemiol 1974;100:350–3.

26 Grayson DA. Confounding confounding. Am J Epidemiol 1987;126:546–53.

27 Weinberg CR. Towards a clearer definition of confounding. Am J Epidemiol 1993;137:1–8.

28 Greenland S, Neutra R. Control of confounding in the assessment of medical technology. Int J Epidemiol 1980;9:361–7.

29 Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran E, Berry D, eds. Statistical Models in Epidemiology: The Environment and Clinical Trials. New York: Springer-Verlag, 1999.

30 Siebert U. Comments from the recipient of the award for outstanding short course: causal inference in decision analysis—DAGs as causal roadmaps. Soc Med Dec Mak Newsl 2005;17:9–10.

31 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10:37–48.

32 Pearl J. Causality. Cambridge, UK: Cambridge University Press, 2000.

33 Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000;11:561–70.

34 Giordano SH, Kuo YF, Duan Z, et al. Limits of observational data in determining outcomes from cancer therapy. Cancer 2008;112:2456–66. Related Articles, Links.

35 Maclure M, Schneeweiss S. Causation of bias: the episcope. Epidemiology 2001;12:114–22.

36 MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. Lancet 2001;357:455–62.

37 Perrio M, Waller PC, Shakir SAW. An analysis of the exclusion criteria used in observational pharmacoepidemiological studies. Pharmacoepidemiol Drug Saf 2006;16:329–36.

38 Schneeweiss S, Patrick AR, Sturmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. Med Care 2007;45(Suppl.):S131–42.

39 Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 2003;158:915–20.

40 Rothman KJ. Epidemiology. An Introduction. New York: Oxford University Press, 2002.

41 Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. Epidemiology 2001;12:682–9.

42 Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. N Engl J Med 1998;338:1516–20.

43 Glynn RJ, Monane M, Gurwitz JH. Choodnovskiy I. Avorn J. Aging, comorbidity, and reduced rates of drug treatment for diabetes mellitus. J Clin Epidemiol 1999;52:781–90.

44 Petri H, Urquhart J. Channeling bias in the interpretation of drug effects. Stat Med 1991;10:577–81.

45 Sturmer T, Rothman KJ, Glynn RJ. Insights into different results from different causal contrasts in the presence of effect-measure modification. Pharmacoepidemiol Drug Saf 2006;15: 698–709.

46 Benner JS, Glynn RJ, Mogun H, et al. Long-term persistence in use of statin therapy in elderly patients. JAMA 2002;288:455–61.

47 Pablos-Mendez A, Barr RG, Shea S. Run-in periods in randomized trials: implications for the application of results in clinical practice. JAMA 1998;279:222–5.

48 Rothwell PM. Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. Lancet 2005;365: 176–86.