A Systematic Statistical Approach to Evaluating Evidence from Observational Studies

David Madigan,[1,2] Paul E. Stang,[2,3] Jesse A. Berlin,[4] Martijn Schuemie,[2,3] J. Marc Overhage,[2,5] Marc A. Suchard,[2,6,7,8] Bill Dumouchel,[2,9] Abraham G. Hartzema,[2,10] and Patrick B. Ryan[2,3]

[1]Department of Statistics, Columbia University, New York, New York 10027; email: david.madigan@columbia.edu

[2]Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, Maryland 20810

[3]Janssen Research and Development LLC, Titusville, New Jersey, 08560

[4]Johnson & Johnson, New Brunswick, New Jersey, 08901; email: pstang@its.jnj.com, jberlin@its.jnj.com, mschuemi@its.jnj.com, pryan4@its.jnj.com

[5]Siemens Health Services, Malvern, Pennsylvania, 19355; email: marc.overhage@siemens.com

[6]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, California, 90095; email: msuchard@ucla.edu

[7]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, 90095

[8]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, California, 90095

[9]Oracle Health Sciences, Burlington, Massachusetts, 01803; email: bill.dumouchel@oracle.com

[10]College of Pharmacy, University of Florida, Gainesville, Florida, 32610; email: hartzema@cop.ufl.edu

Annu. Rev. Stat. Appl. 2014. 1:11–39

First published online as a Review in Advance on November 20, 2013

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-022513-115645

Keywords

pharmacovigilance, epidemiology, data interpretation, statistical, electronic heath records, observational studies

Abstract

Threats to the validity of observational studies on the effects of interventions raise questions about the appropriate role of such studies in decision making. Nonetheless, scholarly journals in fields such as medicine, education, and the social sciences feature many such studies, often with limited exploration of these threats, and the lay press is rife with news stories based on these studies. Consumers of these studies rely on the expertise of the study authors to conduct appropriate analyses, and on the thoroughness of the scientific peer-review process to check the validity, but the introspective and ad hoc nature of the design of these analyses appears to elude any meaningful objective assessment of their performance. Here, we review some of the challenges encountered in observational studies and review an alternative, data-driven approach to observational study design, execution, and analysis. Although much work remains, we believe this research direction shows promise.

1. INTRODUCTION

Consider the following article that recently appeared in the *New England Journal of Medicine* concerning the drug azithromycin and the risk of cardiovascular death (Ray et al. 2012). The paper concludes that, relative to an alternative antibiotic, amoxicillin, "azithromycin was associated with an increased risk of cardiovascular death (hazard ratio, 2.49; 95% CI, 1.38 to 4.50; P = 0.002)" (p. 1881). The authors conducted their study in a database derived from patients enrolled in the Tennessee Medicaid program, identifying 347,795 people with prescriptions for azithromycin and 1,248,672 people with prescriptions for amoxicillin.

Patients, providers, payers, and regulators rely on hundreds or even thousands of observational studies like this to make critical decisions. Nearly 80,000 observational studies were published in the decade 1990–2000 (Naik 2012). In the following decade, the number of studies grew to more than 260,000. Because Ray et al.'s (2012) study was not a randomized clinical trial, two challenges arise. First, the estimated rate ratio may be biased, which is a concern about all observational studies and one that commands most of the discussion section of each paper. That is, this type of analysis could systematically produce rate ratio estimates that are on average too high or too low. Second, the confidence interval might be badly calibrated. That is, the 95% confidence intervals produced by this type of analysis might, for instance, contain the true treatment effect only 50% of the time, instead of the expected 95%. For the example above, a debiased estimate may be 1.0 or 5.0 instead of 2.49. Perhaps the true 95% confidence interval should be 0.50–5.00 instead of 1.38–4.50. What would have happened if different investigators had studied the same question in a different database or applied a different analysis method to the same data? How is this information supposed to contribute to our evidence base? The epidemiologic and statistical literatures acknowledge that observational studies can produce biased and miscalibrated estimates but provide little guidance on how to quantify the extent of the problem. One exception is the highly cited work of Ioannidis (2005), but even this effort at quantification draws mainly on hypothetical scenarios rather than empirical investigation.

This paper reviews the ongoing work of the Observational Medical Outcomes Partnership (OMOP; **http://omop.org**) that attempts to shed light on some of these questions (Stang et al. 2010). Our work to date has focused on four primary objectives. First, we have attempted to establish the operating characteristics of current standard observational study methods (Ryan et al. 2012). Specifically, we have characterized bias, coverage of confidence intervals, and ability to discriminate between positive and negative controls for thousands of implementations of epidemiologic designs across 10 databases for hundreds of drug-outcome pairs. We have collaborated with investigators in a network of European databases who have conducted similar methodological research and have replicated the OMOP experiment (Schuemie et al. 2012). Our results suggest that bias is a significant problem in many contexts, and that statistical measurements, such as confidence intervals and *p*-values, are substantially invalid, empirically confirming Ioannidis's findings. Second, we have documented the substantial variability that results when identical analyses are conducted against different databases or when specific decisions within an analysis, such as time at risk or confounding adjustment strategy, are modified. Third, we have developed data-driven approaches to remove bias and provide confidence intervals and *p*-values with close to nominal operating characteristics (e.g., 95% confidence intervals that contain the true effect size 95% of the time). Fourth, we have developed an approach to observational study design that yields known operating characteristics.

We view our work as taking early steps toward a rigorous, well-characterized, evidence-based approach to estimating effects in observational studies. Much work remains to be done. We believe the use of observational healthcare data could be transformed from episodic investigations based on

presumed expert opinion and anecdote into a comprehensive system that can proactively explore, monitor, and evaluate human disease, health service utilization, and the effects of all medical products across a large array of health outcomes of interest (HOI), in/near real time.

Our work to date focuses specifically on drug safety, an issue of particular current concern. The US Food and Drug Administration (FDA) Amendments Act of 2007 required the establishment of an active postmarket risk identification and analysis system with access to patient-level observational data from 100 million lives by 2012 (**http://www.fda.gov/Safety/FDAsSentinelInitiative/default.htm**). However, we believe our work is applicable to the many other applications of observational studies to estimate the effects of interventions, including the emerging interest in comparative effectiveness research.

Section 2 reviews the literature to highlight challenges in observational data analysis. Section 3 reviews the results of a subset of OMOP experiments that consider method performance for four specific outcomes in five large-scale observational databases. Section 4 describes an approach to calibrating the statistical outputs of an observational study. Specifically, we describe an approach for calculating calibrated *p*-values that provide desired false positive rates. Section 5 sets forth a recipe for observational studies that delivers known performance characteristics. We conclude with a discussion.

2. CHALLENGES IN OBSERVATIONAL ANALYSIS

Prior to regulatory approval, while a drug is in development, randomized clinical trials represent the primary sources of safety information. Such experiments are generally regarded as the highest level of evidence, leading to an unbiased estimate of the average treatment effect (Atkins et al. 2004). Unfortunately, most trials suffer from insufficient sample size and lack of applicability to reliably estimate the risk of many potential safety concerns for the target population (Berlin et al. 2008, Waller & Evans 2003). Even if one leverages meta-analytic tools, rare side effects, long-term outcomes (both positive and negative), and effects in patients with comorbidities may still be unknown when a product is approved because of the relatively small size and short duration of clinical trials. For products intended to treat chronic, non-life-threatening conditions that occur in large populations, the International Conference for Harmonisation (Azoulay et al. 2012) recommends a baseline safety database that involves at least 1,500 patients on average with at least a 6-month exposure time to reliably (i.e., 95% of the time) identify events happening at the 1% level (US Food Drug Admin. 1999). In other words, events that occur less frequently than 1 in 100 patients are not expected to be detected under this recommendation.

Observational studies represent an alternative approach to evaluating drug safety questions and can give us the necessary information about drug effects to support clinical decision making. Observational studies provide empirical investigations of exposures and their effects, but differ from experiments in that assignment of treatment to subjects is not controlled (Rosenbaum 2002). Observational studies can be based on many forms of epidemiologic investigation, using a variety of methods for data collection, applying alternative study designs, and employing a range of analysis strategies (Hartzema et al. 1999, 2008; Jewell 2004; Rothman 2002; Rothman et al. 2008; Szklo & Nieto 2007). These studies can range from population-based cohort studies with prospective data collection to targeted disease registries to retrospective case-control studies.

One type of resource that has provided fertile ground for epidemiologic investigation has been observational healthcare databases. Administrative claims and electronic health record (EHR) databases have been actively used in pharmacoepidemiology for more than 30 years (Strom 2005) but have seen increased use in the past decade owing to greater availability at lower costs and technological advances that made computational processing on large-scale data more feasible.

Many such databases contain large numbers of patients that make possible the examination of rare events and specific subpopulations that previously could not be studied with sufficient power (Rockhill et al. 1998, Rodriguez et al. 2001). Because the data reflect healthcare activity within a real-world population, they offer the potential to complement clinical trial results. Long-term longitudinal capture of data in these sources can also enable studies that monitor the performance of risk management programs or other interventions over time (Weatherby et al. 2002).

Administrative claims databases have been the most actively used observational healthcare data source. These databases typically capture data elements used within the reimbursement process, as providers of healthcare services (e.g., physicians, pharmacies, hospitals, and laboratories) must submit encounter information to enable payment for their services (Hennessy 2006). The submitted information commonly includes pharmacy claims for dispensing of prescription drugs (e.g., the drug dispensed, the dispensing date, and the number of days of drug supply) and medical claims (inpatient and outpatient) that detail the dates and types of services rendered. Medical claims typically contain diagnosis codes used to justify reimbursement for the services provided. Information on over-the-counter drug use and in-hospital medication is usually unavailable, and the patient's compliance with the prescription is generally unknown (Suissa & Garbe 2007).

EHRs generally contain data captured at the point of care, with the intention of supporting the process of clinical care as well as justifying reimbursement and providing data for quality measurement. A patient record may include demographics (birth date, gender, and race), height and weight, and family and medical history. Many EHR systems support provider entry of diagnoses, signs, and symptoms and also capture other clinical observations, such as vital signs, laboratory results, and imaging reports. Beyond those, EHRs may often contain findings of physical examinations and the results of diagnostic tests (Schneeweiss & Avorn 2005). EHR systems usually also have the capability to record other important health status indications, such as alcohol use and smoking status (Lewis & Brensinger 2004), but the data may be missing in many patient records (Hennessy 2006). As a result of discontinuous care within the US healthcare system, a patient may have multiple EHRs scattered across the providers the individual has seen, but rarely are those records integrated, nor can they usually be linked, so each EHR record reflects a different and incomplete perspective of that person's healthcare experience. Recent efforts to advance health information exchange aim to reduce this fragmentation.

Neither administrative claims nor EHRs represent the ideal information required to assess a particular effect. For example, diagnoses recorded on medical claims are used to support justification for the payment for a given visit or procedure; a given diagnosis could represent the condition that the procedure was used to rule out or could be an administrative artifact (e.g., the code used by a medical coder to maximize the reimbursement amount). Some diagnosis codes have been studied through source record verification and have demonstrated adequate performance characteristics (Donahue et al. 1997, García Rodríguez & Pérez Gutthann 1998, Hennessy et al. 2009, Lee et al. 2005, Miller et al. 2008, Pladevall et al. 1996, So et al. 2006, Tunstall-Pedoe 1997, Varas-Lorenzo et al. 2008, Wahl et al. 2010, Wilchesky et al. 2004), whereas other conditions and systems provide less certainty (Harrold et al. 2007, Leonard et al. 2008, Lewis et al. 2007, Strom 2001). Limitations exist in EHR systems as well, in which, apart from concerns about incomplete capture, data may be artificially manipulated to serve clinical care (e.g., an incorrect diagnosis recorded to justify a desired medical procedure). Most systems have insufficient processes to evaluate data quality a priori, requiring intensive work on the part of the researcher to prepare the data for analysis (Hennessy et al. 2007). To estimate potential drug exposures, for example, researchers can make inferences in administrative claims sources based on pharmacy dispensing records, whereas inferences for EHR systems rely on patient self-report and physician prescribing orders

(Hennessy 2006). Neither approach reflects the timing, dose, or duration of drug ingested, so assumptions are required in interpretation of all study results.

The principal concern for all observational studies, which is of particular relevance in observational database evaluation, is the potential for bias. Schneeweiss & Avorn (2005) illustrated some of the potential sources of bias that are introduced throughout the data capture process for both administrative claims and EHRs. An observational study is biased if the treated and control groups differ prior to treatment in ways that can influence the outcome under study (Rosenbaum 2002). Several forms of bias can arise in the design and conduct of an observational study. In the context of drug safety analyses, one of the most challenging issues is confounding by indication, i.e., a situation in which the indication for the medical product is also an independent risk factor for the outcome (Walker 1996). Therefore, a medical product can spuriously appear to be associated with the outcome when no appropriate control for the underlying condition exists, and confounding may persist despite advanced methods for adjustment (Bosco et al. 2010). For example, proton pump inhibitors might produce an apparent increase in risk of gastrointestinal bleeding. This apparent increase could arise because that class of drugs is used to treat symptoms that might also be indicative of such bleeding. A predisposition for healthcare utilization can also produce confounding, perhaps because of functional status, or of access due to proximity, economic, and institutional factors (Brookhart et al. 2010b). An additional concern is immortal time bias, whereby outcomes are not observable within the defined time at risk (Rothman & Suissa 2008; Suissa 2007, 2008).

Several strategies exist for reducing the effects of bias within observational database studies. These include design-level considerations and analysis approaches. Multiple study design approaches have been proposed for observational investigations, including cohort, case control, and self-controlled case series (SCCS), each with its own approach to address confounding. Self-controlled designs aim to address the threat of between-person confounding by comparing exposed and unexposed time at the individual level (Whitaker et al. 2006). Confounders (e.g., sex) that do not change over time within a person are inherently controlled in such designs. Cohort designs compare outcome rates across populations, so they must control confounding by measuring and adjusting for confounding factors that vary among patients. Some believe each design may have potential applications for examining specific types of associations based on the attributes of the exposure and outcome (e.g., certain designs are presumed to be appropriate for short-term effects) (Gagne et al. 2012).

Cohort studies, perhaps the most commonly used design, allow many possible approaches to address confounding. One design strategy is to impose restrictions on the selected sample to increase validity, potentially at the expense of precision. These restrictions are quite analogous to those employed in clinical trials and include ensuring that only incident drug users are studied; the restrictions also ensure similar comparison groups, patients without contraindications, and comparable adherence, as demonstrated by Schneeweiss et al. (2007), who showed how bias was reduced at each stage of restriction using statin and 1-year mortality as an example.

The restriction to incident users deserves special attention, as implementation of a new-user design can eliminate prevalent user bias (Cadarette et al. 2009, Ray 2003, Schneeweiss 2010). Within a new-user design framework, measures of association focus on events occurring after the first initiation of treatment, thus allowing a more direct comparison with an analogous group using an alternative treatment. The design can be logically extended to study drug switching and add-on therapies, as long as incident use of the target drug is preserved (Schneeweiss 2010).

Comparator selection is also an important design consideration to reduce confounding by indication. The comparator definition should ideally yield patients in the same health circumstance as those eligible to be new users of the target medication. Frequently, when assessing a drug safety issue, the comparator is chosen to represent the standard of care that would have been provided to

that patient had the person not been prescribed the target drug, such that relative effect estimates represent risk above and beyond what the patient could otherwise expect. However, a comparator may also be selected specifically to address a question about difference in risk stemming from the underlying biological mechanism (for example, choosing a comparator drug with the same indication). A challenge in comparator selection arises either when no standard of care exists or when significant channeling bias to a particular drug class is present. For example, this bias might occur when a particular class of drugs is reserved for the most severely ill patients, who might be at increased risk for an adverse event because of the increased severity of disease. In this regard, evaluation studies can be highly sensitive to the comparator selected, and a criticism of these studies is often the subjective nature by which the comparator was selected.

Once a design is established, researchers can further reduce bias through analysis strategies, such as matching, stratification, and statistical adjustment. Variables commonly considered for adjustment are those for which the distribution at baseline differs between the exposed and unexposed populations, or those known to potentially influence treatment decisions. To produce confounding, these variables also need to be associated with outcome occurrence; they may include patient demographics (such as age, gender, and race) or patient comorbidities (expressed either as a set of binary classifiers of specific diseases or as a composite index of comorbidity). One commonly used measure is the Charlson index (Bravo et al. 2002; Charlson et al. 1987, 1994; Cleves et al. 1997; D'Hoore et al. 1993, 1996; Li et al. 2008; Needham et al. 2005; Quan et al. 2005; Southern et al. 2004; Zhang et al. 1999), which was originally developed to predict mortality but has also been shown to be related to healthcare expenditures (Farley et al. 2006). Adjustment for a comorbidity index is useful for exploratory data analysis (Schneeweiss et al. 2001) but may not suffice to address all sources of confounding. Additional variables often cited include prior use of medications and markers for health service utilization, such as number of outpatient visits and inpatient stays. The specific definitions and applications of covariates are highly variable across drug safety evaluation studies. Covariate selection can influence the magnitude of effect measures, regardless of the modeling approach undertaken, particularly if effect modification exists (Lunt et al. 2009).

Once variables are identified, one can control for them through direct matching or stratification, whereby the target and comparator groups are logically divided by the attributes of the covariates. However, in a multivariable context, the data may be too sparse to provide adequate sample size to allow matching on all covariates or to provide subpopulations within each covariate-defined stratum (i.e., there may be empty cells defined by combinations of covariates). A popular tool to overcome this limitation is propensity score analysis (Rosenbaum 2002, Rubin 1997).

As with other approaches, the propensity score model is only as good as the covariates selected to provide the adjustment. A propensity score is a single metric that is intended to account for all of the explanatory variables that predict who will receive treatment. Propensity scores generally balance observed confounders but do not necessarily produce balance on factors not incorporated into the model. Such imbalances represent a particular problem for the analysis of databases in which many important covariates, such as smoking status, alcohol consumption, body mass index, and lifestyle and cultural attitudes regarding health, are not captured. Sturmer et al. (2007) demonstrated that further adjustment could be achieved by conducting supplemental validation studies to collect additional information on previously unmeasured confounders. Schneeweiss et al. (2005) showed how unmeasured confounders biased estimates of COX-2 inhibitors and myocardial infarction. Seeger et al. (2003, 2005, 2007) highlighted how a model without the appropriate variables included could yield a biased estimate in a case study that explored association of statin therapy and myocardial infarction. Strategies for automated selection of large sets of covariates have been proposed as potential solutions to reduce the possibility of

missing an empiric confounder (Schneeweiss et al. 2009). Sensitivity analysis has been proposed as an additional approach to assess the potential consequences of unobserved confounding (Schneeweiss 2006) but is unfortunately rarely reported in published studies. For example, Rosenbaum (2002) posits the existence of a latent confounder and explores the magnitude of the confounding that would be required to explain away the observed effect.

Instrumental variable (IV) analysis presents another potential solution to adjusting for confounding through control of a factor that is related to exposure but unrelated to outcome (Brookhart et al. 2010a, Hogan & Lancaster 2004, Schneeweiss 2007). Several studies have shown how IV analysis can reduce bias (Dudl et al. 2009; Rassen et al. 2009, 2010; Schneeweiss et al. 2008). A challenge in IV analysis is identifying a covariate that satisfies the criteria of an IV, particularly with regard to having no association with the outcome. For active surveillance, in which one may explore multiple outcomes for a given exposure, the selection of a common IV becomes even harder.

One consideration for all statistical adjustment techniques in drug safety evaluation studies is the danger of the statistical adjustment itself introducing bias. Statistical control can sometimes either increase bias or decrease precision without affecting bias and can thereby produce less reliable effect estimates (Schisterman et al. 2009). For example, bias can also be induced if an analysis improperly stratifies on a collider variable (Cole et al. 2010), that is, a variable that is itself directly influenced by two other variables. As a result, care is necessary in any evaluation study to develop a parsimonious model that achieves an appropriate balance between bias and variance. Although researchers have made substantial progress to establish theoretical and conceptual arguments for design and analysis considerations in observational data analysis, very little empirical evidence exists to support best practice and determine how observational analyses should be properly interpreted when evaluating the potential effects of medical products.

3. EVALUATING THE PERFORMANCE OF METHODS: THE OBSERVATIONAL MEDICAL OUTCOMES PARTNERSHIP EXPERIMENT

We have conducted a large-scale observational data experiment that seeks to empirically establish the operating characteristics of many standard epidemiologic methods. Specifically, we created a reference set of 399 product-outcome pairs, each classified as either a positive control (i.e., the product increases the risk of the outcome) or a negative control (i.e., the product neither increases nor decreases the risk of the outcome) (Ryan et al. 2013). Across five databases, we assess how well hundreds of different analytic methods can (*a*) discriminate between the positive controls and the negative controls and (*b*) estimate the true relative risk for the negative controls.

3.1. Data and Test Cases

The five databases included in the work presented here are from Truven Health Analytics (formerly the health business of Thomson Reuters), specifically, from MarketScan® Lab Supplement (MSLR; 1.2 million people), Medicare Supplemental Beneficiaries (MDCR; 4.6 million people), Multi-State Medicaid (MDCD; 10.8 million people), Commercial Claims and Encounters (CCAE; 46.5 million people), and the Quintiles Practice Research Database (GE Centricity EHR; 11.2 million people). Quintiles is an EHR database, whereas the other four databases contain administrative claims data. **Table 1** provides further details about each of the databases. All databases were transformed into a common data model using a standardized vocabulary, such that analyses could be consistently applied across the different sources (Overhage et al. 2012).

Table 1 Data sources used in the experiment

Data	Name	Description	Population	Observation time	Drugs	Conditions	Procedures	Observations
CCAE	MarketScan Commercial Claims and Encounters	Represents privately insured population and captures administrative claims with patient-level deidentified data from in/outpatient visits and pharmacy claims of multiple insurance plans	Total: 46.5 million 49% male Mean age 31.4 (18.1)	Patient years: 97.6 million 2003–2009	Records: 1,030.6 million NDC from pharmacy dispensing claims HCPCS/CPT/ ICD9P procedures from in/outpatient medical claims	Records: 1,257.5 million ICD9 from in/outpatient medical claims	Records: 1,979.1 million HCPCS/CPT/ ICD9P procedures from in/outpatient medical claims	Not available
MDCD	MarketScan Multi-State Medicaid	Contains administrative claims data for Medicaid enrollees from multiple states, including inpatient, outpatient, and pharmacy services	Total: 10.8 million 42% male Mean age 21.3 (21.5)	Patient years: 20.7 million 2002–2007	Records: 360.2 million NDC from pharmacy dispensing claims HCPCS/CPT/ ICD9P procedures from in/outpatient medical claims	Records: 552.8 million ICD9 from in/outpatient medical claims	Records: 557.7 million HCPCS/CPT/ ICD9P procedures from in/outpatient medical claims	Not available
MDCR	MarketScan Medicare Supplemental Beneficiaries	Captures administrative claims for retirees with Medicare supplemental insurance paid by employers, including services provided under Medicare-covered payment, employer-paid portion, and any out-of-pocket expenses	Total: 4.6 million 44% male Mean age 73.5 (8.0)	Patient years: 13.4 million 2003–2009	Records: 400.9 million NDC from pharmacy dispensing claims HCPCS/CPT/ ICD9P procedures from in/outpatient medical claims	Records: 404.9 million ICD9 from in/outpatient medical claims	Records: 478.3 million HCPCS/CPT/ ICD9P procedures from in/outpatient medical claims	Not available

MSLR	MarketScan Lab Supplemental	Represents privately insured population that has at least one recorded laboratory value, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results	Total: 1.2 million 35% male Mean age 37.6 (17.7)	Patient years: 2.2 million 2003–2007	Records: 37.6 million NDC from pharmacy dispensing claims HCPCS/CPT/ ICD9P procedures from in/outpatient medical claims	Records: 49.5 million ICD9 from in/outpatient medical claims	Records: 68.5 million HCPCS/CPT/ ICD9P procedures from in/outpatient medical claims	Records: 41.8 million LOINC from outpatient laboratory services
GE	GE Centricity	Derived from data pooled by providers into a data warehouse in a HIPAA-compliant manner using GE Centricity Office (an ambulatory electronic health record)	Total: 11.2 million 42% male Mean age 39.6 (22.0)	Patient years: 22.4 million 1996–2008	Records: 182.6 million GPI from medication history and prescriptions written	Records: 66.1 million ICD9 from problem list	Records: 110.6 million CPT from procedure list	Records: 1,121.1 million LOINC for laboratory values, SNOMED for chief complaints, signs, and symptoms

Patient years represent total observed patient time in each database.

Abbreviations: CCAE, MarketScan Commercial Claims and Encounters; GE, GE Centricity; HCPCS/CPT/ICD9P, Healthcare Common Procedure Coding System/Current Procedural Terminology/International Classification of Diseases; HIPAA, Health Insurance Portability and Accountability Act; LOINC, Logical Observation Identifiers Names and Codes; MDCD, MarketScan Multi-State Medicaid; MDCR, MarketScan Medicare Supplemental Beneficiaries; MSLR, MarketScan Lab Supplemental; NDC, National Drug Code.

We created a large series of positive and negative controls, known drug-outcome associations for four HOI: acute kidney injury, acute liver injury, acute myocardial infarction, and upper gastrointestinal bleeding [definitions are described in Hansen (2013) and available at **http://omop.org/HOI**]. These HOI represent four of the most important drug safety outcomes considered for a risk identification system (Trifirò et al. 2009). Their importance lies in their frequency of occurrence, in their clinical impact, or in both. For each of these outcomes, drug-outcome pairs (Ryan et al. 2013) were classified as positive controls (active ingredients with evidence to suspect a positive association with the outcome) or negative controls (active ingredients with no evidence to expect a causal effect with the outcome) based on the following criteria.

Positive Controls:

- The event is listed in the boxed warning or warnings/precautions section of an active FDA Structured Product Label.
- The drug is listed as a "causative agent" in Tisdale & Miller (2010).
- The literature review identified no adequately powered studies that refuted evidence of effect.

Negative Controls:

- The event is not listed anywhere in any section of an active FDA Structured Product Label.
- The drug not listed as a causative agent in Tisdale & Miller (2010).
- The literature review identified no adequately powered studies with evidence of potential positive association.

The test cases include 165 positive controls and 234 negative controls (Ryan et al. 2013). The article describes the full set of test cases and provides a more detailed description of the manner in which we constructed the set. Note that for the positive controls, we have not attempted to characterize the true effect size, other than hypothesizing that this effect is positive.

3.2. Methods Evaluated

We sought to establish the operating characteristics of seven methods [new-user cohort, case control, SCCS, self-controlled cohort (SCC) (observational screening), disproportionality analysis (DP), information component temporal pattern discovery (ICTPD), and longitudinal gamma Poisson shrinker (LGPS)]. **Table 2** provides a description of each method, as well as the specific implementation choices considered within each method. We believe these methods include all commonly used approaches in observational studies of healthcare.

For each analytic method and combination of analytic design choices, we generated estimated relative risks and associated standard errors for all 399 drug-outcome test cases. The estimates and associated standard errors for all of the analyses are available for download at **http://omop.org/Research**. For every database, we considered only those drug-outcome pairs with sufficient power to detect a hypothetical relative risk of 1.25, based on the age-by-gender-stratified drug and outcome prevalence estimates (i.e., the power calculations were based on marginal distributions, not on observed associations).

3.3. Metrics

To gain insight into the ability of a method to distinguish between positive and negative controls, we used the effect estimates to compute the area under the receiver operating characteristic curve (AUC) (Fawcett 2006), a measure of predictive accuracy: An AUC of 1 indicates a perfect prediction of which test cases are positive and which are not. An AUC of 0.5 is equivalent to random guessing.

Table 2 Methods and design choices used in the experiment

Method	Method description	Analytic design choices (optimal setting for MDCR–acute myocardial infarction in italics)
Self-controlled cohort (SCC) as implemented in the observational screening package	This is an extension of a traditional cohort epidemiology design in which the rate of adverse drug events can be compared across groups of patients exposed to different medications, allowing comparisons within a cohort population, between treatments, as well as relative to the overall population at large.	**Exposures to include**: all occurrences, first occurrence **Outcomes to include**: first occurrence, all occurrences **Time at risk**: length of exposure + 30 days, 30 days from exposure start, all time post–exposure start **Include index date in time at risk**: no, yes **Control period**: length of exposure + 30 days, 30 days prior to exposure start, 180 days prior to exposure start, 365 days prior to exposure start, all time prior to exposure start **Include index date in control period**: no, yes **Combinations to be tested**: 126
Self-controlled case series (SCCS)	The method estimates the association between a transient exposure and adverse event using only cases; no separate controls are required because each case acts as its own control.	**Outcomes to include**: all occurrences, first occurrence **Prior distribution**: normal, Laplace **Variance of the prior**: determined through cross-validation, predefined at 0.01, predefined at 0.1, predefined at 1, predefined at 10 **Time at risk**: all time post–exposure start, length of exposure, length of exposure + 30 days, 30 days from exposure start **Include index date in time at risk**: yes, no **Apply multivariate adjustment on all drugs**: no, yes **Required observation time**: none, 180 days **Combinations to be tested**: 560
Case control	The program applies a case-control surveillance design to estimate odds ratios for drug-condition effects, in which cases are matched to controls by age, sex, location, and race.	**Controls per case**: up to 10 controls per case, up to 100 controls per case **Required observation time prior to outcome**: 30 days, 180 days **Time at risk**: length of exposure + 30 days, length of exposure, 30 days from exposure start, all time post–exposure start **Include index date in time at risk**: no, yes **Case-control matching strategy**: age, sex, and visit (within 180 days); age, sex, and visit (within 30 days); age and sex **Nesting within indicated population**: no, yes **Exposures to include**: first occurrence, all occurrences **Metric**: odds ratio with Mantel-Haenszel adjustment by age and gender, unadjusted odds ratio **Combinations to be tested**: 384
Information component temporal pattern discovery (ICTPD)	This is a novel method for event history data, focusing explicitly on the detailed temporal relationship between pairs of events. The proposed measure contrasts the observed-to-expected ratio in a period of interest with that in a predefined control period.	**Control period**: −180 days to −1 day before exposure start, −1,080 days to −361 days before exposure start, −30 days to −1 day before exposure start, −810 days to −361 days before exposure start **Time at risk**: 360 days from exposure start, 30 days from exposure start, 60 days from exposure start **Use control period in expected calculation**: yes, no **Use 1 month prior to exposure in expected calculation**: no, yes **Use 1 day prior to exposure in expected calculation**: no, yes **Combinations to be tested**: 42

(*Continued*)

Table 2 (*Continued*)

Method	Method description	Analytic design choices (optimal setting for MDCR–acute myocardial infarction in italics)
New-user cohort	This implementation of the inception cohort design applies various approaches for propensity score adjustment to balance baseline covariates and model to estimate drug-related effects.	**Required observation time prior to exposure**: 180 days, none **Nesting within indicated population**: no, yes **Comparator population**: patients with a diagnosis for the indication of the target drug and at least one exposure to a drug known to be unassociated with the outcome; patients with exposure to most prevalent comparator drug that shares the same indication as the target drug but is not in the same pharmacologic class; patients with exposure to any comparator drug that shares the same indication as the target drug but is not in the same pharmacologic class; patients with a diagnosis for the indication of the target drug **Time at risk**: length of exposure + 30 days, 30 days from exposure start, all time post–exposure start **Propensity score covariate-selection strategy**: Bayesian logistic regression using all available covariates, high-dimensional propensity score covariate-selection algorithm by Schneeweiss et al. (2009), exposure-specific covariate-selection algorithm identified by Brookhart et al. (2006), no covariate adjustment **Covariate eligibility window**: 180 days prior to exposure, 30 days prior to exposure, all time prior to exposure, none **Dimensions to include as potential covariates**: drugs, conditions, and procedures; drugs only; drugs and conditions; none **Additional covariates to include in the propensity score model**: age, sex, index year, Charlson index, number of drugs, number of visits, and number of procedures; age and sex; none **Covariate-selection algorithm additional parameters**: BLR: normal prior distribution with variance = 1, Laplace prior distribution with variance = 1; High-dimensional propensity scoring (HDPS): 100 top confounders from among 200 most prevalent covariates in each dimension that occur in at least 100 persons, 500 top confounders from among 500 most prevalent covariates in each dimension that occur in at least 100 persons **Propensity score trimming**: none, trim lower 5% from the comparator group and the upper 5% from the target group **Metric**: propensity score adjustment using propensity score as continuous variable in logistic regression outcome model, propensity score adjustment using 5 strata as indicator variables in logistic regression outcome model, propensity score adjustment using 20 strata as indicator variables in logistic regression outcome model, propensity score stratification using Mantel-Haenszel adjustment over 5 strata, propensity score stratification using Mantel-Haenszel adjustment over 20 strata, unadjusted odds ratio from univariate logistic regression predicting outcome from exposure **Combinations to be tested**: 126

(*Continued*)

Table 2 (*Continued*)

Method	Method description	Analytic design choices (optimal setting for MDCR–acute myocardial infarction in italics)
Disproportionality analysis (DP)	Methods adapted from data mining of spontaneous adverse event reports, in which drug-condition pairs are identified if they co-occur disproportionately more frequently than would be expected if the drug and condition were independent.	**Outcomes to include**: first occurrence, all occurrences **Strategy to stratify data**: classify drug-outcome co-occurrences as exposed/unexposed and with/without outcome **Metric**: proportional reporting ratio (PRR), information component (BCPNN/IC), multi-item gamma Poisson shrinker **Stratify by age**: yes, no **Stratify by gender**: yes, no **Stratify by year**: no, yes **Time at risk**: length of exposure + 30 days, length of exposure + 60 days, 30 days from exposure start, all time post–exposure start **Combinations to be tested**: 48
Longitudinal gamma Poisson shrinker (LGPS)	LGPS applies Bayesian shrinkage to an estimated incidence rate ratio to compare the exposed population with the general population, and LEOPARD aims to detect and discard associations due to protopathic bias.	**Metric**: incidence rate ratio with Mantel-Haenszel adjustment over age-by-gender strata, LGPS **Exposures to include**: all occurrences, first occurrence **Time at risk**: length of exposure, length of exposure + 30 days **Required observation time prior to exposure**: 365 days, none **Apply LEOPARD filtering for protopathic bias**: yes, no **Combinations to be tested**: 32

Often not only are we interested in whether there is an effect, but also we would like to know the magnitude of the effect. However, to evaluate the accuracy of the effect size estimates for a particular analytic method, we must know the true effect size. This true effect size is never known, so we restrict our analysis to the negative controls, for which we assume that the true log relative risk is zero. Using the negative controls in real data, we compute bias, the average difference between the observed log relative risk and zero. An unbiased estimator would yield a bias of zero. The mean squared error (MSE) (Pladevall et al. 1996) is the average squared difference between the log relative risk and zero. Because zero is the true log relative risk, smaller MSEs are desirable. Coverage probability is the fraction of the 95% intervals that include zero. In the case of an unbiased estimator with valid confidence interval estimation, we would expect the coverage probability to be 95%.

3.4. Results of the Performance Evaluation Experiment

Table 3 presents the analytic method that provided the best AUC for each outcome-database combination as well as the associated AUC value. Each six-digit code specifies a particular set of design choices (for details, see the OMOP 2011-2012 Experiment Method Reference spreadsheet available at **http://omop.org/Research**). For every database-outcome combination, self-controlled methods (SCC, SCCS, and ICTPD) provide the optimal performance, with AUCs ranging from a low of 0.77 for acute liver injury in the MDCD database to 1.00 for acute kidney injury in the MSLR database. In general, AUCs are highest for acute kidney injury and lowest for acute liver injury, with acute myocardial infarction and gastrointestinal bleeding in between. Performance across the five data sources is similar despite their substantial differences in patient populations (**Table 1**).

Table 3 AUC (area under the receiver operating characteristic curve) optimal analytic method for each database and outcome

Data source	Acute kidney injury	Acute liver injury	Acute myocardial infarction	Upper gastrointestinal bleeding
MDCR	OS: 401002 (0.92) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: all occurrences Time at risk: length of exposure + 30 days Include index date in time at risk: no Control period: length of exposure + 30 days Include index date in control period: no	OS: 401002 (0.76) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: all occurrences Time at risk: length of exposure + 30 days Include index date in time at risk: no Control period: length of exposure + 30 days Include index date in control period: no	OS: 407002 (0.84) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: first occurrence Time at risk: length of exposure + 30 days Include index date in time at risk: no Control period: length of exposure + 30 days Include index date in control period: no	OS: 402002 (0.86) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: all occurrences Time at risk: length of exposure + 30 days Include index date in time at risk: yes Control period: length of exposure + 30 days Include index date in control period: yes
CCAE	OS: 404002 (0.89) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: all occurrences Time at risk: length of exposure + 30 days Include index date in time at risk: no Control period: length of exposure + 30 days Include index date in control period: yes	OS: 403002 (0.79) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: all occurrences Time at risk: length of exposure + 30 days Include index date in time at risk: yes Control period: length of exposure + 30 days Include index date in control period: no	OS: 408013 (0.85) Study design: self-controlled cohort Exposures to include: first occurrence Outcomes to include: first occurrence after exposure Time at risk: all time post–exposure start Include index date in time at risk: no Control period: all time prior to exposure start Include index date in control period: no	SCCS: 1931010 (0.82) Outcomes to include: all occurrences Prior distribution: normal Variance of the prior: determined through cross-validation Time at risk: all time post–exposure start Include index date in time at risk: no Apply multivariate adjustment on all drugs: no Required observation time: 180 days
MDCD	OS: 408013 (0.82) Study design: self-controlled cohort Exposures to include: first occurrence Outcomes to include: first occurrence after exposure Time at risk: all time post–exposure start Include index date in time at risk: no Control period: all time prior to exposure start Include index date in control period: no	OS: 409013 (0.77) Study design: self-controlled cohort Exposures to include: first occurrence Outcomes to include: first occurrence Time at risk: all time post–exposure start Include index date in time at risk: no Control period: all time prior to exposure start Include index date in control period: no	OS: 407004 (0.80) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: first occurrence Time at risk: length of exposure + 30 days Include index date in time at risk: no Control period: 365 days prior to exposure start Include index date in control period: no	OS: 401004 (0.87) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: all occurrences Time at risk: length of exposure + 30 days Include index date in time at risk: no Control period: 365 days prior to exposure start Include index date in control period: no

(*Continued*)

Table 3 (*Continued*)

Data source	Acute kidney injury	Acute liver injury	Acute myocardial infarction	Upper gastrointestinal bleeding
MSLR	SCCS: 1907010 (1.00) Outcomes to include: all occurrences Prior distribution: normal Variance of the prior: determined through cross-validation Time at risk: all time post–exposure start Include index date in time at risk: no Apply multivariate adjustment on all drugs: yes Required observation time: none	OS: 406002 (0.84) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: first occurrence after exposure Time at risk: length of exposure + 30 days Include index date in time at risk: no Control period: length of exposure + 30 days Include index date in control period: no	OS: 403002 (0.80) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: all occurrences Time at risk: length of exposure + 30 days Include index date in time at risk: yes Control period: length of exposure + 30 days Include index date in control period: no	OS: 403002 (0.83) Study design: self-controlled cohort Exposures to include: all occurrences Outcomes to include: all occurrences Time at risk: length of exposure + 30 days Include index date in time at risk: yes Control period: length of exposure + 30 days Include index date in control period: no
GE	SCCS: 1949010 (0.94) Outcomes to include: all occurrences Prior distribution: normal Variance of the prior: determined through cross-validation Time at risk: 30 days from exposure start Include index date in time at risk: yes Apply multivariate adjustment on all drugs: yes Required observation time: 180 days	OS: 409002 (0.77) Study design: self-controlled cohort Exposures to include: first occurrence Outcomes to include: first occurrence Time at risk: length of exposure + 30 days Include index date in time at risk: no Control period: length of exposure + 30 days Include index date in control period: no	ICTPD: 3016001 (0.89) Control period: −1,080 days to −361 days before exposure start Time at risk: 60 days from exposure start Use control period in expected calculation: yes Use 1 month prior to exposure in expected calculation: yes Use 1 day prior to exposure in expected calculation: no	ICTPD: 3034001 (0.89) Control period: −810 days to −361 days before exposure start Time at risk: 60 days from exposure start Use control period in expected calculation: yes Use 1 month prior to exposure in expected calculation: no Use 1 day prior to exposure in expected calculation: yes

Abbreviations: CCAE, MarketScan Commercial Claims and Encounters; GE, GE Centricity; ICTPD, information component temporal pattern discovery; MDCD, MarketScan Multi-State Medicaid; MDCR, MarketScan Medicare Supplemental Beneficiaries; MSLR, MarketScan Lab Supplemental; OS, observational screening; SCCS, self-controlled case series.

Figure 1 presents the AUC value for all analytic methods, broken down by database and method. Several findings emerge from **Figure 1**:

- The case-control method, LGPS, and DP consistently underperform other methods, often yielding AUCs close to 0.5.
- Within each method, the specific design choices that correspond to the global optimum generally perform well for all outcomes and databases. Consider, for example, the SCC design; with the exception of acute myocardial infarction in MDCD, performance of the database-outcome optimum design choices does not exceed the global optimum by more than 0.10 in AUC.
- The design choices within each method affect performance significantly. For the majority of drug-outcome-method triples, there are design choices that yield AUC values at or close

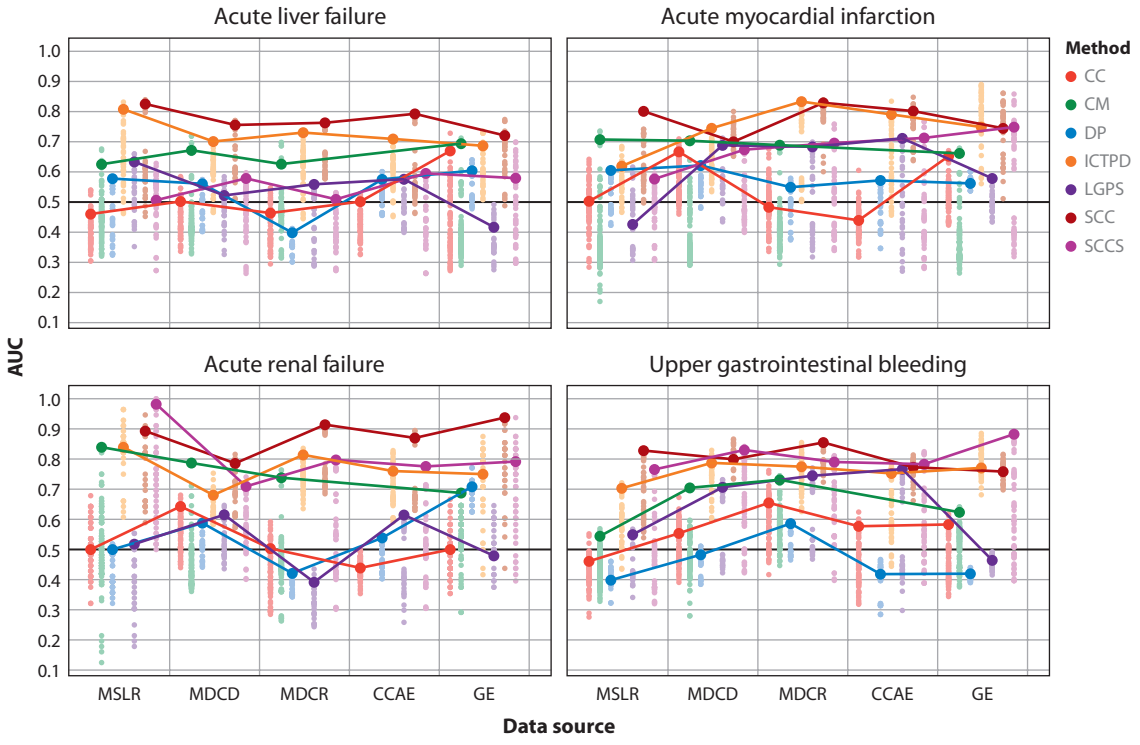


Figure 1

Area under the receiver operating characteristic curve (AUC) values for all analytic methods, broken down by database and method. The solid lines represent the AUCs for the set of design choices within each method that provided the best performance on average across all outcomes and databases, a specific version of global optimum. Abbreviations: CC, case control; CCAE, MarketScan Commercial Claims and Encounters; CM, cohort method; DP, disproportionality analysis; GE, GE Centricity; ICTPD, information component temporal pattern discovery; LGPS, longitudinal gamma Poisson shrinker; MDCD, MarketScan Multi-State Medicaid; MDCR, MarketScan Medicare Supplemental Beneficiaries; MSLR, MarketScan Lab Supplemental; SCC, self-controlled cohort; SCCS, self-controlled case series.

to 0.5, despite the existence of design choices with quite high AUC values for the same drug-outcome pair.

Table 4 considers bias, MSE, and 95% confidence interval coverage for database-outcome optimal analytic methods. Because the true relative risks are unavailable for the positive controls, the table just draws on the negative control test cases. **Table 4** shows that, in our experiments, the case-control, SCC, and LGPS methods generally yield positively biased effect estimates, whereas the cohort method generally yields negatively biased estimates. The SCCS method yields estimates that are close to unbiased. All three self-controlled methods produce smaller MSEs than the other methods, with SCCS being especially close to zero. No method provides coverage probabilities that are close to the nominal 95%. On average, coverage probabilities for the cohort, disproportionality, ICTPD, LGPS, and SCC methods are all below 50%. Average coverage for the case-control method is 63%, whereas for SCCS, the average coverage is 76%.

Figure 2 presents the point estimates for the negative controls across all analysis methods. The positive bias of the case-control, SCC, and LGPS methods reveals itself, as does the negative bias of the cohort method. The smaller MSE associated with SCCS is also apparent.

3.5. Discussion of the Experimental Results

Although no method resulted in perfect discrimination, many methods were substantially better than random guessing. Optimum AUCs ranging from 0.76 to 0.94 seem promising and are at least as good as the predictive accuracy that is typically observed for diagnostic tests used in routine clinical practice. These results suggest that observational data can play an important role in the assessment of the effects of medical products, but no single analysis can provide definitive evidence.

Self-controlled designs (SCC, ICTPD, and SCCS) performed well across all 20 database-outcome scenarios. They generally outperformed case-control and new-user cohort designs in terms of predictive accuracy and did not exhibit notably different error distributions or bias.

All methods have poor coverage probability. This issue is related both to the bias in the point estimate (the difference between the point estimate and the true effect) and to underestimation of the standard error when generating the confidence intervals, and it may be systemic, plaguing the entire enterprise of observational database analysis. Part of the problem is understandable: Confidence intervals convey only error due to sampling variability around an unbiased estimator, but these databases provide extremely large samples that result in modest sampling variability. However, the true culprit in errant observational database studies is systematic error due to, for example, residual confounding and misclassification, which are not accounted for in traditional calculations. Furthermore, unlike sampling variability, systematic error does not diminish as sample size increases. The implication of this fact is that traditional statistical methods for controlling bias, e.g., covariate adjustment or matching, are failing, in these examples, to correct for bias.

Our results suggest that performance improvements result from customizing analyses to databases. Different databases represent different source populations and different data capture processes, and thus some sources might be better at addressing specific questions. Each database exhibits unique limitations that could affect performance. For example, owing to high turnover, payer-based claims data may provide shorter longitudinal capture, whereas outpatient EHR systems may have more incomplete capture during the observation period.

We have not established the generalizability of these findings. That we see different analyses yield the highest predictive accuracy for different database-outcome pairs suggests caution is necessary when projecting these results to other databases or to other outcomes. Further experiments are needed to determine the degree to which results can be generalized across outcomes. A possible direction is to conduct similar experiments for an additional 19 outcomes identified by the Exploring and Understanding Adverse Drug Reactions project (**http://www.euadr-project.org**) as high-priority safety issues (Trifirò et al. 2009).

Our analytic methods do not have access to the labels for the test cases, that is, whether a particular test case was a positive control or a negative control. Thus, our results are not optimistically biased in the sense they would be in a machine learning experiment that reported performance on training data. However, our results pertain to the specific 399 test cases that we studied, and generalizing to future test cases requires, at the very least, an exchangeability assumption that may or may not be reasonable.

4. CALIBRATING RESULTS

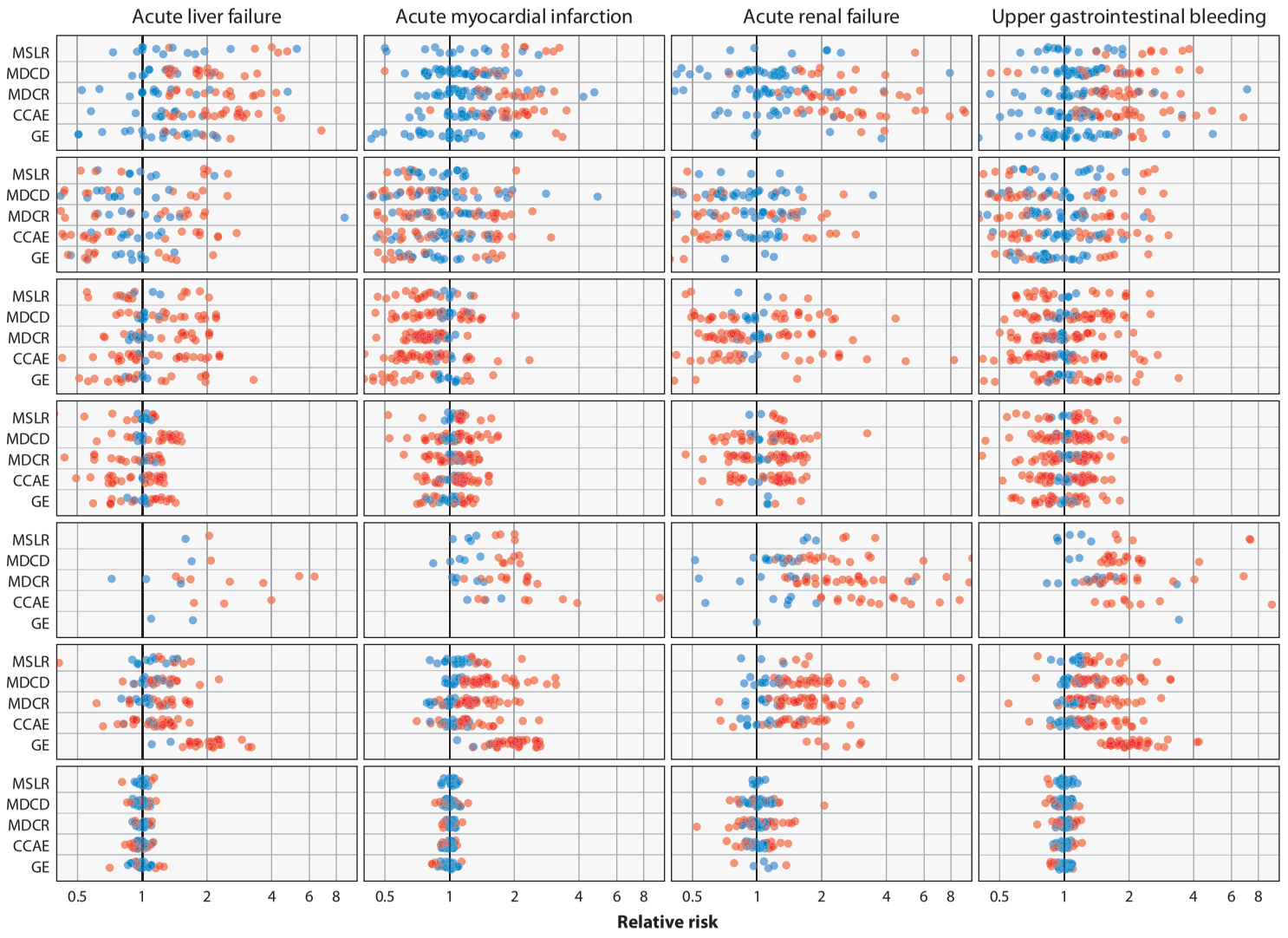
The experimental results in Section 2 demonstrate the possibility that typical observational studies may not account for all sources of bias leading to, for example, confidence intervals with poor coverage properties. Similarly, *p*-values from observational studies may mislead. For example, hypothesis tests with a nominal 5% α level may yield false positive rates that substantially depart from 5%. In this section, we explore this issue and present empirically calibrated *p*-values that

Table 4 Bias, mean square error (MSE), and coverage for the optimal analytic method for each database and outcome

Data Source	Acute liver failure			Acute myocardial infarction			Acute renal failure			Upper gastrointestinal bleeding			
	Bias	MSE	Coverage	Bias	MSE	Coverage	Bias	MSE	Coverage	Bias	MSE	Coverage	Method
MSLR	0.25	0.72	0.72	0.18	0.93	0.68	0.26	0.72	0.86	0.16	0.40	0.65	CC
MDCD	0.20	0.31	0.36	0.04	0.13	0.84	0.05	0.39	0.73	0.13	0.26	0.59	
MDCR	0.21	0.54	0.50	0.15	0.34	0.67	0.19	0.52	0.63	0.13	0.43	0.52	
CCAE	0.28	0.62	0.31	0.15	0.29	0.50	0.33	1.30	0.38	0.15	0.45	0.51	
GE	0.14	0.38	0.73	0.01	0.19	0.95	0.33	0.92	0.67	0.07	0.27	0.87	
MSLR	−0.09	0.71	0.39	−0.09	0.17	0.73	−0.28	0.84	0.71	−0.04	0.42	0.54	CM
MDCD	−0.15	0.55	0.46	−0.20	3.51	0.60	−0.21	2.95	0.69	−0.45	8.92	0.45	
MDCR	−0.16	4.35	0.40	0.02	1.16	0.40	−0.03	0.22	0.45	−0.05	0.31	0.50	
CCAE	−0.05	0.32	0.28	−0.06	0.65	0.51	−0.06	0.32	0.48	−0.14	2.44	0.40	
GE	−0.13	0.44	0.43	−0.30	4.53	0.53	−0.20	0.42	0.50	−0.21	4.07	0.61	
MSLR	0.01	0.14	0.22	−0.10	0.12	0.27	−0.13	0.46	0.29	0.00	0.17	0.17	DP
MDCD	0.06	0.26	0.28	−0.09	0.34	0.26	−0.13	0.86	0.19	0.01	0.20	0.18	
MDCR	0.07	0.14	0.32	−0.10	0.08	0.10	−0.04	0.15	0.20	−0.03	0.08	0.21	
CCAE	0.04	0.22	0.09	−0.14	0.30	0.09	−0.06	0.76	0.09	−0.02	0.18	0.21	
GE	−0.00	0.23	0.19	−0.24	0.64	0.23	−0.34	0.98	0.00	0.01	0.28	0.26	
MSLR	−0.03	0.08	0.44	0.02	0.05	0.32	0.07	0.04	0.29	0.00	0.07	0.17	ICTPD
MDCD	0.06	0.07	0.21	0.02	0.05	0.24	0.06	0.12	0.12	0.01	0.07	0.12	
MDCR	−0.04	0.11	0.25	0.02	0.03	0.19	0.04	0.10	0.10	0.01	0.07	0.15	
CCAE	−0.02	0.07	0.06	0.03	0.04	0.13	0.04	0.08	0.09	0.02	0.06	0.15	
GE	−0.00	0.05	0.27	−0.00	0.03	0.49	0.04	0.08	0.50	−0.02	0.07	0.21	

MSLR	0.26	0.36	0.50	0.17	0.21	0.50	0.35	0.72	0.50	0.27	0.92	0.50	LGPS
MDCD	0.27	0.41	0.50	0.18	0.26	0.42	0.24	0.63	0.39	0.31	0.86	0.10	
MDCR	0.31	0.96	0.30	0.21	0.33	0.31	0.38	1.20	0.19	0.23	0.48	0.36	
CCAE	0.41	1.00	0.00	0.36	1.05	0.30	0.48	2.02	0.27	0.42	1.41	0.00	
GE	0.14	0.15	1.00				−0.00	0.00	1.00	0.53	1.51	1.00	
MSLR	0.06	0.12	0.56	0.07	0.07	0.68	0.12	0.13	0.43	0.11	0.12	0.46	SCC
MDCD	0.08	0.07	0.34	0.15	0.20	0.28	0.21	0.41	0.26	0.14	0.18	0.24	
MDCR	0.05	0.07	0.39	0.07	0.08	0.33	0.18	0.28	0.20	0.11	0.14	0.21	
CCAE	0.05	0.05	0.16	0.06	0.07	0.35	0.12	0.18	0.29	0.09	0.10	0.32	
GE	0.30	0.52	0.08	0.28	0.48	0.08	0.37	0.77	0.00	0.30	0.62	0.00	
MSLR	0.00	0.01	0.83	0.01	0.00	0.95	0.00	0.00	1.00	−0.00	0.01	0.92	SCCS
MDCD	−0.00	0.00	0.69	0.00	0.00	0.78	−0.00	0.05	0.71	−0.00	0.00	0.80	
MDCR	−0.00	0.00	0.71	0.00	0.00	0.81	0.00	0.04	0.49	0.00	0.01	0.64	
CCAE	−0.01	0.00	0.53	−0.00	0.00	0.78	0.00	0.02	0.59	0.00	0.00	0.77	
GE	−0.00	0.01	0.77	−0.01	0.01	0.87	0.03	0.04	0.67	−0.01	0.00	0.85	

Abbreviations: CC, case control; CCAE, MarketScan Commercial Claims and Encounters; CM, cohort method; DP, disproportionality analysis; GE, GE Centricity; ICTPD, information component temporal pattern discovery; LGPS, longitudinal gamma Poisson shrinker; MDCD, MarketScan Multi-State Medicaid; MDCR, MarketScan Medicare Supplemental Beneficiaries; MSLR, MarketScan Lab Supplemental; SCC, self-controlled cohort; SCCS, self-controlled case series.



may support more appropriate inferences and decisions. The core idea is to derive an empirical null distribution using the negative control test cases described in Section 4.1.

4.1. Empirical Null Distributions

To demonstrate the empirical calibration, we selected three drug safety study exemplars from the literature representing a case-control, a cohort, and an SCCS design. We attempted to replicate these studies as best as we could but used databases different from those used in the original articles. The cohort study investigated the relationship between isoniazid and acute liver injury, whereas the case-control and SCCS studies investigated the association between sertraline and upper gastrointestinal bleeding. Our replications produced very similar effect estimates, each falling within the 95% confidence interval reported in the original article.

Next, we applied the same three designs to sets of negative controls: drugs that are not believed to cause the outcome of interest. **Figure 3** shows the estimated odds ratios and incidence rate ratios, together with associated 95% confidence intervals. As is evident in **Figure 3**, traditional significance testing fails to capture the diversity in estimates that exists when the null hypothesis is true. Although all the featured drug-outcome pairs are negative controls (i.e., assumed to have a true odds ratio or rate ratio of 1), a large fraction of the null hypotheses are rejected. We would expect only 5% of negative controls to have $p < 0.05$. However, in **Figure 3*a*** (cohort method), 17 of the 30 negative controls (57%) are either significantly protective or significantly harmful. In **Figure 3*b*** (case-control method), 33 of 46 negative controls (72%) are significantly harmful. Similarly, in **Figure 3*c*** (SCCS method), 33 of 45 negative controls (73%) are significantly harmful, although not the same 33 as in **Figure 3*b***.

These numbers cast doubts on any observational study that would claim statistical significance. Consider, for example, the odds ratio of 2.4 that we found for sertraline using the case-control method; we see in **Figure 3*b*** that many of the negative controls have similar or even higher odds ratios. The estimate for sertraline was highly significant ($p < 0.001$), meaning the null hypothesis can be rejected based on the usual theoretical model, i.e., assuming a null distribution centered at a relative rate of 1.0. However, based on the empirical distribution of negative controls, we can argue that we should not reject the null hypothesis so readily.

4.2. Calibration

Using the empirical distributions of negative controls, we can compute a better estimate of the probability that a value at least as extreme as a certain effect estimate could have been observed under the null hypothesis (Schuemie et al. 2013).

Figure 4 shows the fraction of negative controls for which the p-value is below α for every level of α, for both the traditional p-value calculation and the calibrated p-value using the empirically established null distribution. For the calibrated p-value, a leave-one-out design was used: For each

Figure 2

Point estimates for the negative controls across all analysis methods. Red dots indicate estimates for which the corresponding 95% confidence interval does not include one, whereas blue dots indicate estimates for which the corresponding 95% confidence interval includes zero (i.e., a relative risk of one on the original scale; in an ideal situation, with unbiased estimators, 95% of all dots should blue). Abbreviations: CC, case control; CCAE, MarketScan Commercial Claims and Encounters; CM, cohort method; DP, disproportionality analysis; GE, GE Centricity; ICTPD, information component temporal pattern discovery; LGPS, longitudinal gamma Poisson shrinker; MDCD, MarketScan Multi-State Medicaid; MDCR, MarketScan Medicare Supplemental Beneficiaries; MSLR, MarketScan Lab Supplemental; SCC, self-controlled cohort; SCCS, self-controlled case series.

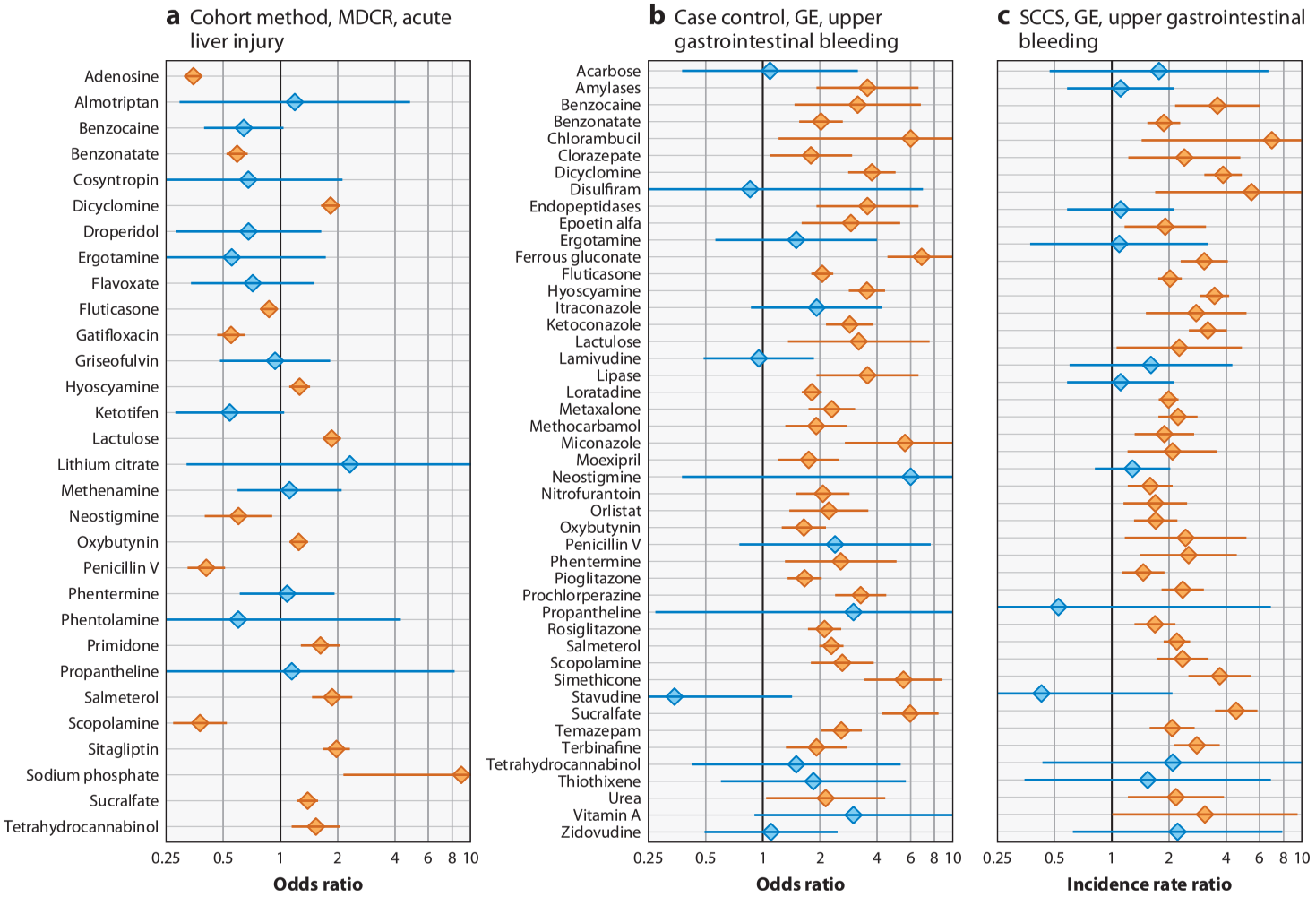


Figure 3

Forest plots of negative controls. Lines show 95% confidence intervals. Orange indicates statistically significant estimates (two-sided $p < 0.05$); blue indicates nonsignificant estimates. Abbreviations: GE, GE Centricity; MDCR, MarketScan Medicare Supplemental Beneficiaries; SCCS, self-controlled case series.

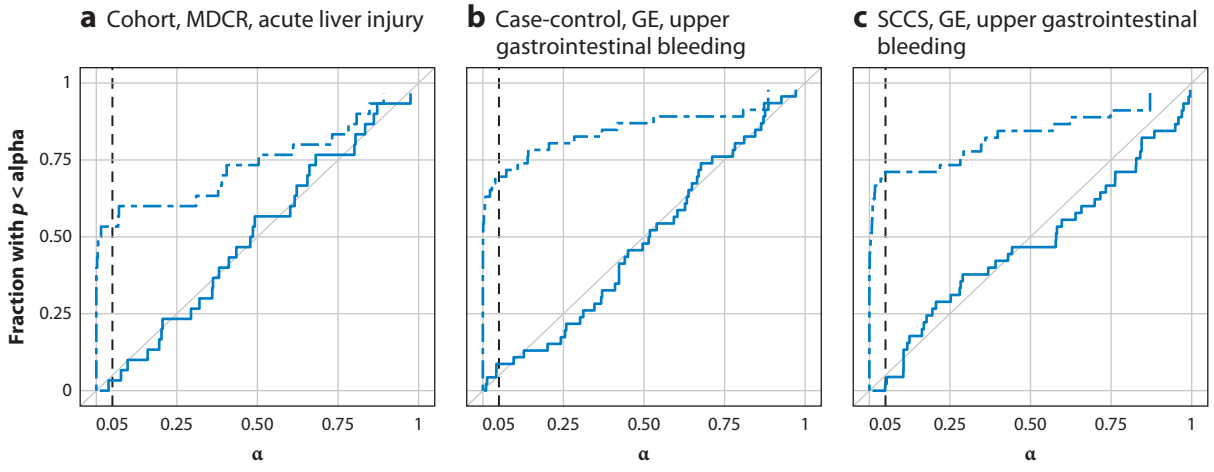


Figure 4

Calibration plots, showing the fraction of negative controls with $p < \alpha$ for different levels of α. Both traditional p-value calculation (*dashed lines*) and p-values using calibration (*solid lines*) are shown. For the calibrated p-value, a leave-one-out design was used. Abbreviations: GE, GE Centricity; MDCR, MarketScan Medicare Supplemental Beneficiaries; SCCS, self-controlled case series.

negative control, the null distribution was estimated using all other negative controls. A well-calibrated p-value calculation should follow the diagonal: For negative controls, the proportion of estimates with $p < \alpha$ should be approximately equal to α. Most significance testing uses an α of 0.05, and we see in **Figure 4** that the calibrated p-value leads to the desired level of rejection of the null hypothesis. For the cohort, case-control, and SCCS methods, the number of significant negative controls after calibration is 2 of 34 (6%), 5 of 46 (11%), and 3 of 46 (5%), respectively.

Applying the calibration to our three example studies, we find that only the cohort study of isoniazid reaches statistical significance ($p = 0.01$). The case-control and SCCS analyses produced p-values of 0.71 and 0.84, respectively. Note that the calibration process could theoretically result in a larger p-value although we have not seen this happen in practice.

4.3. Visualization of the Calibration

Figure 5 shows a graphical representation of the calibration. By plotting the effect estimate on the x-axis and the standard error of the estimate on the y-axis, we can visualize the area where the traditional p-value is smaller than 0.05 (the gray area below the dashed line) and where the calibrated p-value is smaller than 0.05 (orange area). Many of the negative controls fall within the gray area, indicating a traditional $p < 0.05$, but only a few fall within the orange area, indicating a calibrated $p < 0.05$.

In **Figure 5*a***, the drug of interest, isoniazid, is clearly separated from the negative controls, and this separation is the reason why we feel confident we can reject the null hypothesis of no effect. In **Figure 5*b*,*c***, the drug of interest, sertraline, is indistinguishable from the negative controls. These studies provide little evidence for rejecting the null hypothesis.

5. A RECIPE FOR OBSERVATIONAL STUDIES

Current strategies for the design of observational studies rely heavily on the expertise of analysts. A process of expert consideration, introspection, anecdote, and discussion leads to a particular

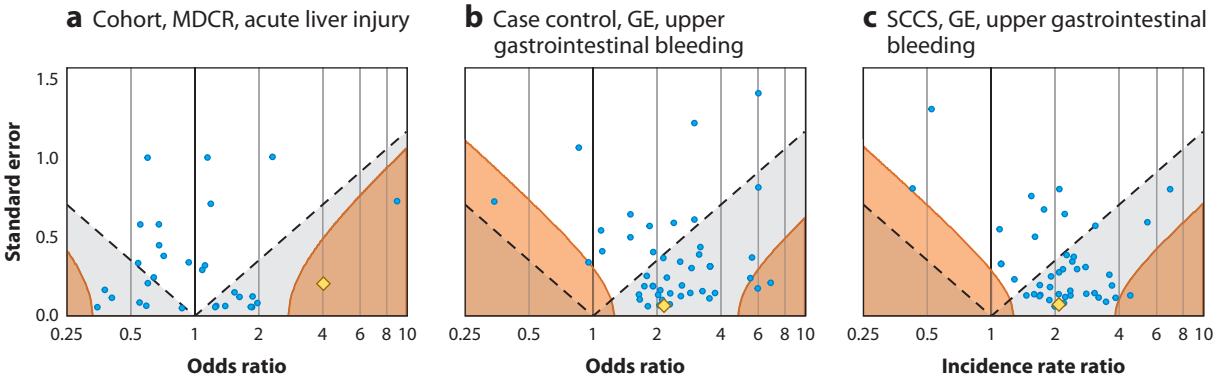


Figure 5

Traditional and calibrated significance testing. Estimates below the dashed line (*gray area*) have $p < 0.05$ using traditional p-value calculation. Estimates in the orange areas have $p < 0.05$ using the calibrated p-value calculation. In each panel, blue dots indicate negative controls, whereas the yellow diamond indicates the particular drug of interest: (*a*) isoniazid and (*b*,*c*) sertraline. Abbreviations: GE, GE Centricity; MDCR, MarketScan Medicare Supplemental Beneficiaries; SCCS, self-controlled case series.

design. The consumer of the resulting analysis must rely on the professional experience and reputation of the analyst to assess the weight of evidence to attach to the study. Because little empirical evidence exists to support this process, the subjective assessment of new results, along with prior beliefs about the reliability of observational studies, dominates the interpretation of observational findings in current practice.

Our work suggests an alternative path that provides a data-driven approach to study design. We found that all methods have error, and the magnitude/direction of error varies by analysis and database. Although some analyses carried substantial information, with AUC > 0.80, considerable bias and large MSE existed, with p-values and confidence intervals far from nominal operating characteristics. Future work can lead to improvements in methods that may have better performance, but our findings suggest that empirical evidence will be required to justify interpreting observational analyses properly. However, we also found that the empirical evidence generated can be used to improve the operating characteristics through calibration. Suppose an analyst wishes to study the association between intervention I and outcome O in observational database D. We tentatively recommend proceeding as follows:

1. Develop a set of test cases for O. That is, identify a set of positive control interventions that are known to be associated with O and a set of negative control interventions that are known to be unassociated with O.
2. Run every possible study design in D to generate design-specific estimates and standard errors for all test cases.
3. Choose the study design that optimizes some desired combination of AUC, MSE, bias, and coverage.
4. Report calibrated p-values (and, in due course, calibrated confidence intervals) for the optimal design along with the performance characteristics associated with the design, as computed in step 3.
5. As a sensitivity exploration, report calibrated confidence intervals and p-values for the other high-performing designs along with their performance characteristics.

Whether such an approach yields better results than the current expert-centric strategy remains unknown. A key limitation of our recipe is that it is specific to the database and to the outcome

only. Future work will consider extensions that also account for characteristics of the target drug.

6. DISCUSSION

We have demonstrated the empirical performance of various analytic methods for observational studies across a range of database-outcome combinations. The ability of an analytic method to discriminate between positive and negative test cases, rather than the more traditional subjective judgment of an expert analyst, dictates the customization. No method resulted in perfect discrimination, but many methods were substantially better than random guessing.

Section 4 reviewed a procedure for calculating calibrated *p*-values for observational studies. We have developed a related procedure for calculating calibrated confidence intervals that involves inverting the hypothesis test associated with the calibrated *p*-value. Thus, for example, the value zero is in the 95% calibrated confidence interval for the true log relative risk if the two-sided calibrated *p*-value exceeds 0.05. For values other than zero, we resort to injecting known relative risks into simulated data (see Murray et al. 2011 for a complete description of the simulation procedure).

We believe that the path forward for observational studies lies not only in more thoughtful and careful study design but also in augmenting current practice by applying a rigorous and empirically-driven systematic approach to study design. This approach considers observational studies as instruments for making measurements. Only by systematically measuring and comparing performance of well-characterized processes can we hope to improve our ability to measure the strength of association between drug exposure and outcome and to distinguish between positive and negative effects. If the process is thoughtful and careful but ad hoc, comparing its application across different problems is impossible, and therefore the process cannot be improved. As Popper (1965) has pointed out, "I knew, of course, the most widely accepted answer to my problem: that science is distinguished from pseudoscience—or from 'metaphysics'—by its empirical method, which is essentially inductive, proceeding from observation or experiment." Applying an empirically based approach to observational analysis offers tremendous potential for meaningfully and reliably contributing to the scientific evidence base needed to support medical practice.

DISCLOSURE STATEMENT

Johnson & Johnson, one of many organizations that provide funding for the Observational Medical Outcomes Partnership (OMOP), is the employer of J.A.B. and P.E.S. A.G.H. and D.M. were paid consultants to OMOP. P.B.R. is an employee of Janssen Research and Development. M.S. received a fellowship from the Office of Medical Policy, Center for Drug Evaluation and Research, US Food and Drug Administration, and has become an employee of Janssen Research and Development since completing the research presented here.

ACKNOWLEDGMENTS

M.A.S. received funding from the Foundation for the National Institutes of Health. This work was also partially supported by National Science Foundation grant IIS1251151.

LITERATURE CITED

Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, et al. 2004. Grading quality of evidence and strength of recommendations. *BMJ* 328:1490

Azoulay L, Yin H, Filion KB, Assayag J, Majdan A, et al. 2012. The use of pioglitazone and the risk of bladder cancer in people with type 2 diabetes: nested case-control study. *BMJ* 344:e3645

Berlin JA, Glasser SC, Ellenberg SS. 2008. Adverse event detection in drug development: recommendations and obligations beyond phase 3. *Am. J. Public Health* 98:1366–71

Bosco JL, Silliman RA, Thwin SS, Geiger AM, Buist DS, et al. 2010. A most stubborn bias: No adjustment method fully resolves confounding by indication in observational studies. *J. Clin. Epidemiol.* 63:64–74

Bravo G, Dubois MF, Hébert R, De Wals P, Messier L. 2002. A prospective evaluation of the Charlson Comorbidity Index for use in long-term care patients. *J. Am. Geriatr. Soc.* 50:740–45

Brookhart MA, Rassen JA, Schneeweiss S. 2010a. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol. Drug Saf.* 19:537–54

Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. 2010b. Confounding control in healthcare database research: challenges and potential approaches. *Med. Care* 48:S114–20

Cadarette SM, Katz JN, Brookhart MA, Sturmer T, Stedman MR, et al. 2009. Comparative gastrointestinal safety of weekly oral bisphosphonates. *Osteoporos. Int.* 20:1735–47

Charlson ME, Pompei P, Ales KL, MacKenzie CR. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* 40:373–83

Charlson ME, Szatrowski TP, Peterson J, Gold J. 1994. Validation of a combined comorbidity index. *J. Clin. Epidemiol.* 47:1245–51

Cleves MA, Sanchez N, Draheim M. 1997. Evaluation of two competing methods for calculating Charlson's comorbidity index when analyzing short-term mortality using administrative data. *J. Clin. Epidemiol.* 50:903–8

Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, et al. 2010. Illustrating bias due to conditioning on a collider. *Int. J. Epidemiol.* 39:417–20

D'Hoore W, Bouckaert A, Tilquin C. 1996. Practical considerations on the use of the Charlson comorbidity index with administrative data bases. *J. Clin. Epidemiol.* 49:1429–33

D'Hoore W, Sicotte C, Tilquin C. 1993. Risk adjustment in outcome assessment: the Charlson comorbidity index. *Methods Inf. Med.* 32:382–87

Donahue JG, Weiss ST, Goetsch MA, Livingston JM, Greineder DK, Platt R. 1997. Assessment of asthma using automated and full-text medical records. *J. Asthma* 34:273–81

Dudl RJ, Wang MC, Wong M, Bellows J. 2009. Preventing myocardial infarction and stroke with a simplified bundle of cardioprotective medications. *Am. J. Manag. Care* 15:e88–94

Farley JF, Harley CR, Devine JW. 2006. A comparison of comorbidity measurements to predict healthcare expenditures. *Am. J. Manag. Care* 12:110–19

Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognit. Lett.* 27:861–74

Gagne JJ, Fireman B, Ryan PB, Maclure M, Gerhard T, et al. 2012. Design considerations in an active medical product safety monitoring system. *Pharmacoepidemiol. Drug Saf.* 21(Suppl. 1):32–40

García Rodríguez LA, Pérez Gutthann S. 1998. Use of the UK General Practice Research Database for pharmacoepidemiology. *Br. J. Clin. Pharmacol.* 45:419–25

Hansen RA, Gray MD, Fox BI, Hollingsworth JC, Gao J, Zeng P. 2013. How well do various health outcome definitions identify appropriate cases in observational studies? *Drug Saf.* 36(Suppl. 1):S27–32

Harrold LR, Saag KG, Yood RA, Mikuls TR, Andrade SE, et al. 2007. Validity of gout diagnoses in administrative data. *Arthritis Rheum.* 57:103–8

Hartzema AG, Porta MS, Tilson HH. 1999. *Pharmacoepidemiology: An Introduction*. Cincinnati, OH: Harvey Whitney Books

Hartzema AG, Tilson HH, Chan KA. 2008. *Pharmacoepidemiology and Therapeutic Risk Management*. Cincinnati, OH: Harvey Whitney Books

Hennessy S. 2006. Use of health care databases in pharmacoepidemiology. *Basic Clin. Pharmacol. Toxicol.* 98:311–13

Hennessy S, Leonard CE, Freeman CP, Deo R, Newcomb C, et al. 2009. Validation of diagnostic codes for outpatient-originating sudden cardiac death and ventricular arrhythmia in Medicaid and Medicare claims data. *Pharmacoepidemiol. Drug Saf.* 19:555–62

Hennessy S, Leonard CE, Palumbo CM, Newcomb C, Bilker WB. 2007. Quality of Medicaid and Medicare data obtained through Centers for Medicare and Medicaid Services (CMS). *Med. Care* 45:1216–20

Hogan JW, Lancaster T. 2004. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Stat. Methods Med. Res.* 13:17–48
Ioannidis JP. 2005. Why most published research findings are false. *PLoS Med.* 2:e124
Jewell N. 2004. *Statistics for Epidemiology*. Boca Raton, FL: Chapman & Hall
Lee DS, Donovan L, Austin PC, Gong Y, Liu PP, et al. 2005. Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. *Med. Care* 43:182–88
Leonard CE, Haynes K, Localio AR, Hennessy S, Tjia J, et al. 2008. Diagnostic E-codes for commonly used, narrow therapeutic index medications poorly predict adverse drug events. *J. Clin. Epidemiol.* 61:561–71
Lewis JD, Brensinger C. 2004. Agreement between GPRD smoking data: a survey of general practitioners and a population-based survey. *Pharmacoepidemiol. Drug Saf.* 13:437–41
Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. 2007. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol. Drug Saf.* 16:393–401
Li B, Evans D, Faris P, Dean S, Quan H. 2008. Risk adjustment performance of Charlson and Elixhauser comorbidities in ICD-9 and ICD-10 administrative databases. *BMC Health Serv. Res.* 8:12
Lunt M, Solomon D, Rothman K, Glynn R, Hyrich K, et al. 2009. Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *Am. J. Epidemiol.* 169:909–17
Miller DR, Oliveria SA, Berlowitz DR, Fincke BG, Stang P, Lillienfeld DE. 2008. Angioedema incidence in US veterans initiating angiotensin-converting enzyme inhibitors. *Hypertension* 51:1624–30
Murray RE, Ryan PB, Reisinger SJ. 2011. Design and validation of a data simulation model for longitudinal healthcare data. *AMIA Annu. Symp. Proc.* 2011:1176–85
Naik G. 2012. Analytical trend troubles scientists. *Wall Street Journal*, May 4
Needham DM, Scales DC, Laupacis A, Pronovost PJ. 2005. A systematic review of the Charlson comorbidity index using Canadian administrative databases: a perspective on risk adjustment in critical care research. *J. Crit. Care* 20:12–19
Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. 2012. Validation of a common data model for active safety surveillance research. *J. Am. Med. Inform. Assoc.* 19:54–60
Pladevall M, Goff DC, Nichaman MZ, Chan F, Ramsey D, et al. 1996. An assessment of the validity of ICD Code 410 to identify hospital admissions for myocardial infarction: the Corpus Christi Heart Project. *Int. J. Epidemiol.* 25:948–52
Popper KR. 1965. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge & Kegan Paul
Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, et al. 2005. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* 43:1130–39
Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. 2009. Instrumental variables I: Instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J. Clin. Epidemiol.* 62:1226–32
Rassen JA, Mittleman MA, Glynn RJ, Brookhart MA, Schneeweiss S. 2010. Safety and effectiveness of bivalirudin in routine care of patients undergoing percutaneous coronary intervention. *Eur. Heart J.* 31:561–72
Ray WA. 2003. Evaluating medication effects outside of clinical trials: new-user designs. *Am. J. Epidemiol.* 158:915–20
Ray WA, Murray KT, Hall K, Arbogast PG, Stein CM. 2012. Azithromycin and the risk of cardiovascular death. *N. Engl. J. Med.* 366:1881–90
Rockhill B, Newman B, Weinberg C. 1998. Use and misuse of population attributable fractions. *Am. J. Public Health* 88:15–19
Rodriguez EM, Staffa JA, Graham DJ. 2001. The role of databases in drug postmarketing surveillance. *Pharmacoepidemiol. Drug Saf.* 10:407–10
Rosenbaum PR. 2002. *Observational Studies*. New York: Springer. 2nd ed.
Rothman KJ. 2002. *Epidemiology: An Introduction*. Oxford: Oxford Univ. Press
Rothman KJ, Greenland S, Lash T. 2008. *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins
Rothman KJ, Suissa S. 2008. Exclusion of immortal person-time. *Pharmacoepidemiol. Drug Saf.* 17:1036
Rubin DB. 1997. Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.* 127:757–63

Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. 2012. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat. Med.* 31:4401–15

Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. 2013. Defining a reference set to support methodological research in drug safety. *Drug Saf.* 36(Suppl. 1):33–47

Schisterman EF, Cole SR, Platt RW. 2009. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 20:488–95

Schneeweiss S. 2006. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol. Drug Saf.* 15:291–303

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

Schneeweiss S. 2010. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol. Drug Saf.* 19:858–68

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

Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. 2005. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. *Epidemiology* 16:17–24

Schneeweiss S, Patrick AR, Sturmer T, Brookhart MA, Avorn J, et al. 2007. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Med. Care* 45:S131–42

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. 2009. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 20:512–22

Schneeweiss S, Seeger JD, Landon J, Walker AM. 2008. Aprotinin during coronary-artery bypass grafting and risk of death. *N. Engl. J. Med.* 358:771–83

Schneeweiss S, Seeger JD, Maclure M, Wang PS, Avorn J, Glynn RJ. 2001. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. *Am. J. Epidemiol.* 154:854–64

Schuemie MJ, Coloma PM, Straatman H, Herings RM, Trifirò G, et al. 2012. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. *Med. Care* 50:890–97

Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. 2013. Interpreting observational studies—why empirical calibration is needed to correct *p*-values. *Stat. Med.* In press. doi: 10.1002/sim.5925

Seeger JD, Kurth T, Walker AM. 2007. Use of propensity score technique to account for exposure-related covariates: an example and lesson. *Med. Care* 45:S143–48

Seeger JD, Walker AM, Williams PL, Saperia GM, Sacks FM. 2003. A propensity score-matched cohort study of the effect of statins, mainly fluvastatin, on the occurrence of acute myocardial infarction. *Am. J. Cardiol.* 92:1447–51

Seeger JD, Williams PL, Walker AM. 2005. An application of propensity score matching using claims data. *Pharmacoepidemiol. Drug Saf.* 14:465–76

So L, Evans D, Quan H. 2006. ICD-10 coding algorithms for defining comorbidities of acute myocardial infarction. *BMC Health Serv. Res.* 6:161

Southern DA, Quan H, Ghali WA. 2004. Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. *Med. Care* 42:355–60

Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, et al. 2010. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann. Intern. Med.* 153:600–6

Strom BL. 2001. Data validity issues in using claims data. *Pharmacoepidemiol. Drug Saf.* 10:389–92

Strom BL. 2005. *Pharmacoepidemiology*. Chichester, UK: Wiley

Sturmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S. 2007. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. *Med. Care* 45:S158–65

Suissa S. 2007. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol. Drug Saf.* 16:241–49

Suissa S. 2008. Immortal time bias in pharmacoepidemiology. *Am. J. Epidemiol.* 167:492–99
Suissa S, Garbe E. 2007. Primer: administrative health databases in observational studies of drug effects—advantages and disadvantages. *Nat. Clin. Pract. Rheumatol.* 3:725–32
Szklo M, Nieto FJ. 2007. *Epidemiology: Beyond the Basics*. Burlington, MA: Jones & Bartlett
Tisdale J, Miller D. 2010. *Drug-Induced Diseases: Prevention, Detection, and Management*. Bethesda, MD: Am. Soc. Health-Syst. Pharm. 2nd ed.
Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, et al. 2009. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol. Drug Saf.* 18:1176–84
Tunstall-Pedoe H. 1997. Validity of ICD code 410 to identify hospital admission for myocardial infarction. *Int. J. Epidemiol.* 26:461–62
US Food Drug Adm. 1999. *Managing the Risks from Medical Product Use: Creating a Risk Management Framework*. Silver Springs, MD: US Food Drug Admin. **http://www.fda.gov/downloads/Safety/SafetyofSpecificProducts/UCM180520.pdf**
Varas-Lorenzo C, Castellsague J, Stang MR, Tomas L, Aguado J, Perez-Gutthann S. 2008. Positive predictive value of ICD-9 codes 410 and 411 in the identification of cases of acute coronary syndromes in the Saskatchewan Hospital automated database. *Pharmacoepidemiol. Drug Saf.* 17:842–52
Wahl PM, Rodgers K, Schneeweiss S, Gage BF, Butler J, et al. 2010. Validation of claims-based diagnostic and procedure codes for cardiovascular and gastrointestinal serious adverse events in a commercially-insured population. *Pharmacoepidemiol. Drug Saf.* 19:596–603
Walker AM. 1996. Confounding by indication. *Epidemiology* 7:335–36
Waller PC, Evans SJ. 2003. A model for the future conduct of pharmacovigilance. *Pharmacoepidemiol. Drug Saf.* 12:17–29
Weatherby LB, Nordstrom BL, Fife D, Walker AM. 2002. The impact of wording in "Dear doctor" letters and in black box labels. *Clin. Pharmacol. Ther.* 72:735–42
Whitaker HJ, Farrington CP, Spiessens B, Musonda P. 2006. Tutorial in biostatistics: the self-controlled case series method. *Stat. Med.* 25:1768–97
Wilchesky M, Tamblyn RM, Huang A. 2004. Validation of diagnostic codes within medical services claims. *J. Clin. Epidemiol.* 57:131–41
Zhang JX, Iwashyna TJ, Christakis NA. 1999. The performance of different lookback periods and sources of information for Charlson comorbidity adjustment in Medicare claims. *Med. Care* 37:1128–39

Annual Review of Statistics and Its Application

Volume 1, 2014

Contents