Drug Saf (2013) 36 (Suppl 1):S143–S158
DOI 10.1007/s40264-013-0108-9

ORIGINAL RESEARCH ARTICLE

A Comparison of the Empirical Performance of Methods for a Risk Identification System

Patrick B. Ryan · Paul E. Stang · J. Marc Overhage · Marc A. Suchard · Abraham G. Hartzema · William DuMouchel · Christian G. Reich · Martijn J. Schuemie · David Madigan

Abstract

Background Observational healthcare data offer the potential to enable identification of risks of medical products, and the medical literature is replete with analyses that aim to accomplish this objective. A number of established analytic methods dominate the literature but their operating characteristics in real-world settings remain unknown.

Objectives To compare the performance of seven methods (new user cohort, case control, self-controlled case series, self-controlled cohort, disproportionality analysis, temporal pattern discovery, and longitudinal gamma poisson shrinker) as tools for risk identification in observational healthcare data.

Research Design The experiment applied each method to 399 drug-outcome scenarios (165 positive controls and 234 negative controls across 4 health outcomes of interest) in 5 real observational databases (4 administrative claims and 1 electronic health record).

Measures Method performance was evaluated through Area Under the receiver operator characteristics Curve (AUC), bias, mean square error, and confidence interval coverage probability.

Results Multiple methods offer strong predictive accuracy, with AUC > 0.70 achievable for all outcomes and databases with more than one analytical approach. Self-controlled methods (self-controlled case series, temporal pattern discovery, self-controlled cohort) had higher predictive accuracy than cohort and case–control methods across all databases and outcomes. Methods differed in the expected value and variance of the error distribution. All methods had lower coverage probability than the expected nominal properties.

Conclusions Observational healthcare data can inform risk identification of medical product effects on acute liver

The OMOP research used data from Truven Health Analytics (formerly the Health Business of Thomson Reuters), and includes MarketScan® Research Databases, represented with MarketScan Lab Supplemental (MSLR, 1.2 m persons), MarketScan Medicare Supplemental Beneficiaries (MDCR, 4.6 m persons), MarketScan Multi-State Medicaid (MDCD, 10.8 m persons), MarketScan Commercial Claims and Encounters (CCAE, 46.5 m persons). Data also provided by Quintiles® Practice Research Database (formerly General Electric's Electronic Health Record, 11.2 m persons) database. GE is an electronic health record database while the other four databases contain administrative claims data.

P. B. Ryan (✉) · P. E. Stang
Janssen Research and Development LLC,
1125 Trenton-Harbourton Road, Room K30205,
PO Box 200, Titusville, NJ 08560, USA
e-mail: ryan@omop.org

J. M. Overhage
Siemens Health Services, Malvern, PA, USA

M. A. Suchard
Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA

M. A. Suchard
Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, USA

A. G. Hartzema
College of Pharmacy, University of Florida,
Gainesville, FL, USA

W. DuMouchel
Oracle Health Sciences, Burlington, MA, USA

C. G. Reich
AstraZeneca, Waltham, MA, USA

injury, acute myocardial infarction, acute renal failure and gastrointestinal bleeding. However, effect estimates from all methods require calibration to address inconsistency in method operating characteristics. Further empirical evaluation is required to gauge the generalizability of these findings to other databases and outcomes.

1 Background

Characterization of risks associated with medical products represents a cornerstone of medical science. Prior to regulatory approval, while a drug is in development, randomized clinical trials (RCTs) represent the primary sources of product safety information. RCTs are often thought of as the gold standard in estimating true causal effects, and evidence from trials is highly regarded when available [1]. Unfortunately, most trials suffer from insufficient sample size and lack of applicability to reliably estimate the risk of other potential safety concerns for the target population [2, 3]. As a result, new evidence about safety is required even after a medical product is approved. Evidence from non-randomized observational studies has become a critical component of the evidence base about the safety profile of a medicine after approval, but the degree to which the evidence from observational analyses is consistent with 'truth' is not well understood.

Increasing attention now focuses on the secondary use of observational healthcare databases (e.g., administrative claims and electronic health records) for risk identification, as these sources reflect "real world" experience of patient populations of interest to research in pharmacoepidemiology, health outcomes and health services research. In this context, analysts use a variety of statistical/epidemiological methods to produce risk estimates and associated statistical artifacts such as confidence intervals and p-values. However method selection and implementation rely on highly subjective judgments and the operating characteristics of the resulting risk estimates remain elusive. By "operating characteristics" we mean, for example, the frequency with which, in practice, a 95 % confidence interval for a risk difference actually contains the true risk difference, or the actual type I error rate associated with procedures that reject the null hypothesis of no effect at the 5 % level. A substantial literature describes an array of potential threats to the validity of observational studies and raises concerns that the operating characteristics may well depart from the nominal values [4].

We have conducted a large-scale observational data experiment that aims to establish the operating characteristics of a number of standard observational analysis methods. Our work focuses specifically on drug safety, an issue of particular current concern. The U.S. Food and Drug Administration (FDA) Amendments Act of 2007 required the establishment of an "active post-market risk identification and analysis system" with access to patient-level observational data from 100 million lives by 2012 [5]. This system, developed under the umbrella of the "Sentinel Initiative," applied to a network of observational healthcare databases will provide another source of evidence to complement existing safety information contributed by preclinical data, clinical trials, spontaneous adverse event reports, registries, and pharmacoepidemiology evaluation studies. The Mini-Sentinel pilot system [6] is already generating risk estimates that inform FDA communications [7]. The Observational Medical Outcomes Partnership (OMOP; http://omop.org) conducts methodological research to support the Sentinel Initiative.

The OMOP research consists of a series of empirical assessments of the performance characteristics of a number of analysis methods conducted across a network of observational data sources. 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). 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. In other papers within this supplement, each method is detailed and the absolute performance is evaluated [8–14]. In this study, we compare the performance across the various methods and highlight trends we observe across the analytical approaches, databases and health outcomes of interest.

2 Methods

2.1 Experiment Design

We conducted the experiment in five observational healthcare databases to allow evaluation of performance across different populations and data capture processes:

M. J. Schuemie
Department of Medical Informatics, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands

D. Madigan
Department of Statistics, Columbia University, New York, NY, USA

P. B. Ryan · P. E. Stang · J. M. Overhage · M. A. Suchard · A. G. Hartzema · W. DuMouchel · C. G. Reich · M. J. Schuemie · D. Madigan
Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, MD, USA

Table 1 Data sources used in the experiment

Abbreviation	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 de-identified data from inpatient and outpatient visits and pharmacy claims of multiple insurance plans	Total: 46.5 m % male: 49 % Mean age [41]: 31.4 (18.1)	Pt-years: 97.6 m 2003–2009	Records: 1,030.6 m NDC from pharmacy dispensing claims HCPCS/CPT/ICD9P procedures from inpatient/outpatient medical claims	Records: 1,257.5 m ICD9 from inpatient/ outpatient medical claims	Records: 1979.1 m HCPCS/CPT/ ICD9P procedures from inpatient/ 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 m % male: 42 % Mean age [41]: 21.3 (21.5)	Pt-years: 20.7 m 2002–2007	Records: 360.2 m NDC from pharmacy dispensing claims HCPCS/CPT/ICD9P procedures from inpatient/outpatient medical claims	Records : 552.8 m ICD9 from inpatient/ outpatient medical claims	Records: 557.7 m HCPCS/CPT/ ICD9P procedures from inpatient/ 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 m % male: 44 % Mean age [41]: 73.5 (8.0)	Pt-years: 13.4 m 2003–2009	Records: 400.9 m NDC from pharmacy dispensing claims HCPCS/CPT/ICD9P procedures from inpatient/outpatient medical claims	Records : 404.9 m ICD9 from inpatient/ outpatient medical claims	Records: 478.3 m HCPCS/CPT/ ICD9P procedures from inpatient/ 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 m % male: 35 % Mean age [41]: 37.6 (17.7)	Pt-years: 2.2 m 2003–2007	Records: 37.6 m NDC from pharmacy dispensing claims HCPCS/CPT/ICD9P procedures from inpatient/outpatient medical claims	Records: 49.5 m ICD9 from inpatient/ outpatient medical claims	Records: 68.5 m HCPCS/CPT/ ICD9P procedures from inpatient/ outpatient medical claims	Records: 41.8 m LOINC from outpatient laboratory services
GE	GE Centricity™	Derived from data pooled by providers using GE Centricity Office (an ambulatory electronic health record) into a data warehouse in a HIPAA-compliant manner	Total: 11.2 m % male: 42 % Mean age [41]: 39.6 (22.0)	Pt-years: 22.4 m 1996–2008	Records: 182.6 m GPI from medication history and prescriptions written	Records: 66.1 m ICD9 from problem list	Records: 110.6 m CPT from procedure list	Records: 1,121.1 m LOINC for laboratory values, SNOMED for chief complaints, signs and symptoms

MarketScan® Lab Supplemental (MSLR, 1.2M persons), MarketScan® Medicare Supplemental Beneficiaries (MDCR, 4.6M persons), MarketScan® Multi-State Medicaid (MDCD, 10.8M persons), MarketScan® Commercial Claims and Encounters (CCAE, 46.5M persons), and the General Electric Centricity™ (GE, 11.2M persons) database. GE is an electronic health record (EHR) database, whereas the other four databases contain administrative claims data. Table 1 provides further details about each of the databases.

We sought to establish the operating characteristics of seven methods (new user cohort [11], case control [9], self-controlled case series [14], self-controlled cohort [12], disproportionality analysis [8], temporal pattern discovery [10], and longitudinal gamma poisson shrinker [13]). Table 2 provides a description of each method, as well as the specific analysis choices considered within each method. Each method represents a different approach to effect estimation with varying strategies for adjusting for potential sources of bias in observational analyses [15]. Self-controlled case series, self-controlled cohort, and temporal pattern discovery can all be classified as self-controlled designs, as each avoids between-person confounding by using the same subjects in both exposed and control groups, albeit at different points in time (i.e., when exposed and not-exposed). The new user cohort and case–control methods attempt to address between-person confounding by balancing external populations and providing relative comparisons. The extent to which methods employ covariate adjustment for fixed and time-dependent confounding varies by method. This paper provides an overall comparison of the results across all analyses within all methods. Additional findings about the absolute performance of each method and the relative considerations of analysis choices within each method are available in the method-specific papers in this Supplement cited above.

For each method and combination of analysis choices within each method we generated estimated relative risks and associated standard errors for 399 drug-outcome test cases. Some analysis choices are similar across methods (ex. time-at-risk), and in those cases, we attempted to evaluate similar analysis choice values across methods. Many analysis choices were specific to each method's implementation (ex. external comparator group selection and propensity score adjustment strategy within cohort method, controls-per-case within the case–control design). The estimates and associated standard errors for all of the analyses are available for download at: http://omop.org/Research.

These test cases include 165 'positive controls'—active ingredients with evidence to suspect a positive association with the outcome—and 234 'negative controls'—active ingredients with no evidence to expect a causal effect with the outcome—and were limited to four outcomes: acute liver injury, acute myocardial infarction, acute renal failure, and upper gastrointestinal bleeding. Ryan et al. [16] describes the full set of test cases and its construction. For every database we only considered those drug-outcome pairs with sufficient power to detect a relative risk of 1.25, based on the age-by-gender-stratified drug and outcome prevalence estimates.

2.2 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 curve (AUC), 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.

Often we are not only interested in whether there is an effect or not, but would also like to know the magnitude of the effect. However, in order 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, and so we restrict our analysis to the negative controls where 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 log relative risk and zero. An unbiased estimator would yield a bias of zero. The "mean square error (MSE)" is the average squared difference between the log relative risk and zero. Since zero is the true relative risk, smaller MSEs are desirable. "Coverage" is the fraction of the 95 % intervals that include zero. In the case of an unbiased estimator with accurate confidence interval estimation we would expect the coverage probability to be 95 %.

3 Results

Table 3 presents the analysis 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 analysis choices within a given method—for details see http://omop.org/Research. For every database-outcome combination, self-controlled methods (SCC, SCCS, and ICTPD) provide the optimal performance with AUC's 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 significant differences (e.g., GE is an

Table 2 Methods and design choices used in the experiment

Method	Analysis choices
Self-controlled cohort (SCC) as implemented in the Observational Screening (OS) package This is an extension of a traditional cohort epidemiology design where the rate of ADEs 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 + 30d, 30d from exposure start, All time post-exposure start **Include index date in time-at-risk:** No, Yes **Control period:** Length of exposure + 30d, 30d prior to exposure start, 180d prior to exposure start, 365d 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, Pre-defined at 0.01, Pre-defined at 0.1, Pre-defined at 1, Pre-defined at 10 **Time-at-risk:** All time post-exposure start, Length of exposure, Length of exposure + 30d, 30d from exposure start **Include index date in time-at-risk:** Yes, No **Apply multivariate adjustment on all drugs:** No, Yes **Required observation time:** None, 180d Combinations to be tested: 560
Case control (CC) The program applies a case–control surveillance design to estimate odds ratios for drug–condition effects, where 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:** 30d, 180d **Time-at-risk:** Length of exposure + 30d, Length of exposure, 30d 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 180d), Age sex and visit (within 30d), 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
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:** −180d to −1d before exposure start, −1,080d to −361d before exposure start, −30d to −1d before exposure start, −810d to −361d before exposure start **Time-at-risk:** 360d from exposure start, 30d from exposure start, 60d from exposure start **Use control period in expected calculation:** Yes, No **Use 1 mo prior to exposure in expected calculation:** No, Yes **Use 1d prior to exposure in expected calculation:** No, Yes Combinations to be tested: 42

Table 2 continued

Method	Analysis choices
New user cohort (CM) 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:** 180d, 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 not associated with the outcome; Patients with exposure to most prevalent comparator drug which shares the same indication as the target drug but is not in the same pharmacologic class; Patients with exposure to any comparator drug which 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 + 30d, 30d 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. Exposure-specific covariate selection algorithm identified by Brookhart et al. No covariate adjustment **Covariate eligibility window:** 180d prior to exposure, 30d 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 include in the propensity score model:** Age and sex and index year and Charlson index and number of drugs and 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; 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

Table 2 continued

Method	Analysis choices
Disproportionality analysis (DP) Methods adapted from data mining of spontaneous adverse event reports, where drug–condition pairs are identified if they co-occur disproportionately more frequently than 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 + 30d, Length of exposure + 60d, 30d 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, Longitudinal Gamma Poisson Shrinker **Exposures to include**: All occurrences, First occurrence **Time-at-risk:** Length of exposure, Length of exposure + 30d **Required observation time prior to exposure:** 365d, None **Apply LEOPARD filtering for protopathic bias:** Yes, No Combinations to be tested: 32

electronic health record database whereas the other four are administrative claims databases).

Figure 1 presents that AUC value for all analyses, again broken down by database and method. The solid lines in Fig. 1 represent the AUCs for the set of analysis choices within each method that provided the best performance on average across all outcomes and databases, a sort of "global optimum." Several findings emerge from Fig. 1:

- The case–control method (CC), LGPS, and disproportionality analyses (DP) consistently underperform other methods often yielding AUC close to 0.5.
- Within each method, the specific analysis choices that correspond to the global optimum generally perform well for all outcomes and databases. Consider, for example, the self-controlled cohort (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.
- The analysis choices within each method impact performance significantly. For the majority of drug-outcome-method triples, there exist analysis choices that yield AUC values at or close to 0.5.

The complete data underlying Fig. 1 are available at: http://omop.org/Research.

Table 4 considers bias, mean square error, and 95 % confidence interval coverage for database-outcome optimal analyses for each method. Since 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 method, self-controlled cohort, and LGPS generally yield positively biased effect estimates where 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 method, disproportionality, temporal pattern discovery, LGPS, and self-controlled cohort 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 the optimum analysis for each method. Red dots indicate estimates where the corresponding 95 % confidence interval does not include one while blue dots indicate estimates where the corresponding 95 % confidence interval does include zero. (Note that in an ideal situation with unbiased estimators, 95 % of all dots should be blue). The positive bias of the case control, self-controlled cohort, and LGPS methods reveals itself, as does the

Table 3 AUC optimal analytic method for each database and outcome

Data source	Acute kidney injury	Acute liver injury	Acute myocardial infarction	Upper gastrointestinal bleed
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 + 30d Include index date in time-at-risk: No Control period: Length of exposure + 30d 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 + 30d Include index date in time-at-risk: No Control period: Length of exposure + 30d 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 + 30d Include index date in time-at-risk: No Control period: Length of exposure + 30d 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 + 30d Include index date in time-at-risk: Yes Control period: Length of exposure + 30d 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 + 30d Include index date in time-at-risk: No Control period: Length of exposure + 30d 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 + 30d Include index date in time-at-risk: Yes Control period: Length of exposure + 30d 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: 180d
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 + 30d Include index date in time-at-risk: No Control period: 365d 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 + 30d Include index date in time-at-risk: No Control period: 365d prior to exposure start Include index date in control period: No

Table 3 continued

Data source	Acute kidney injury	Acute liver injury	Acute myocardial infarction	Upper gastrointestinal bleed
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 + 30d Include index date in time-at-risk: No Control period: Length of exposure + 30d 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 + 30d Include index date in time-at-risk: Yes Control period: Length of exposure + 30d 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 + 30d Include index date in time-at-risk: Yes Control period: Length of exposure + 30d 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: 30d from exposure start Include index date in time-at-risk: Yes Apply multivariate adjustment on all drugs: Yes Required observation time: 180d	**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 + 30d Include index date in time-at-risk: No Control period: Length of exposure + 30d Include index date in control period: No	**ICTPD: 3016001** (0.89) Control period: -1080d to -361d before exposure start Time-at-risk: 60d from exposure start Use control period in expected calculation: Yes Use 1mo prior to exposure in expected calculation: Yes Use 1d prior to exposure in expected calculation: No	**ICTPD: 3034001** (0.89) Control period: -810d to -361d before exposure start Time-at-risk: 60d from exposure start Use control period in expected calculation: Yes Use 1mo prior to exposure in expected calculation: No Use 1d prior to exposure in expected calculation: Yes

MDCR MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *MDCD* MarketScan Multi-state Medicaid, *MSLR* MarketScan Lab Supplemental, *GE* GE Centricity

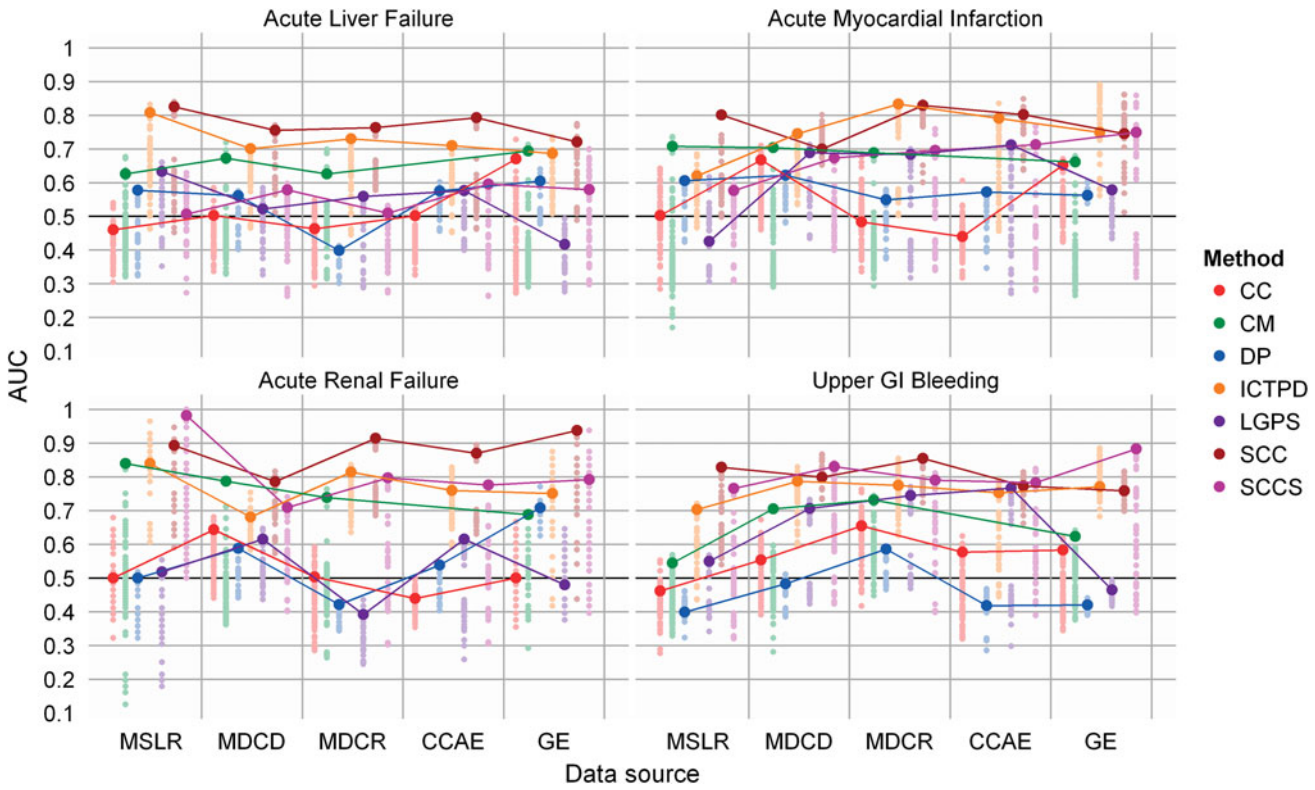

Fig. 1 AUC values for all analytic methods, again broken down by database and method. The *solid lines* in figure represent the AUCs for the set of design choices within each method that provided the best performance on average across all outcomes and databases. Note that the *left*-to-*right* order of the methods in this graph is the same as the *top*-to-*down* order in the legend. *MSLR* MarketScan Lab Supplemental, ***MDCD*** MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity, *AUC* area under the ROC curve, *CC* case control, *CM* cohort method, *DP* disproportionality method, *ICTPD* information component temporal pattern discovery, *LGPS* Longitudinal Gamma Poisson Shrinker, *SCC* self-controlled cohort, *SCCS* self-controlled case series

negative bias of the cohort method. The smaller MSE associated with self-controlled case series is also apparent.

4 Discussion

We have demonstrated the empirical performance of various analytic methods for observational studies customized to the context of each database-outcome combination. The ability of a method to discriminate between positive and negative test cases dictates the customization rather than the more traditional subject judgment of an expert analyst. No analysis resulted in perfect discrimination, but many analysis choices within some methods were substantially better than random guessing. Optimum AUCs ranging from 0.76–0.94 seem promising, and are higher than the predictive accuracy that is typically observed for diagnostic tests used in routine clinical practice. This suggests 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 methods (self-controlled cohort, temporal pattern discovery, self-controlled case series) performed well across all 20 database-outcome scenarios. They generally outperformed new user cohort and case–control designs, in terms of predictive accuracy, and do not exhibit notably different error distributions or bias. Further research is certainly required to confirm this finding, and to explore whether alternative approaches within the new user cohort and/or case–control designs could yield better performance than was observed here. Given that self-controlled designs are not applied as commonly as cohort designs, and are generally perceived to only apply to a narrow set of circumstances [17–20], our findings suggest that a reexamination of some long-held ideas about self-controlled methods is in order.

All methods have poor coverage probability. This is related both to the variability in the error distribution (the difference between the point estimate and the true effect), as well as underestimation of the standard error when generating the confidence intervals. This may be a systemic issue plaguing the entire enterprise of observational

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

	Acute liver failure			Acute myocardial infarction			Acute renal failure			Upper GI bleeding			Method
	Bias	MSE	Coverage	Bias	MSE	Coverage	Bias	MSE	Coverage	Bias	MSE	Coverage	
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	

MDCR MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *MDCD* MarketScan Multi-state Medicaid, *MSLR* MarketScan Lab Supplemental, *GE* GE Centricity

database analysis. Part of this is understandable: confidence intervals only convey 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 confounding, misclassification, etc., which are not represented in traditional calculations. Furthermore, unlike sampling variability, systematic error does not diminish as sample size increases.

These results are applicable to understanding the performance of observational analyses. That includes whether the analysis is performed for one drug-outcome pair at a time (as some refer to as 'signal refinement' or 'signal evaluation') or whether the analysis is performed across multiple drug-outcome pairs simultaneous (as some refer to as 'signal detection'). In all of these use cases, the objective remains the same: to produce a credible estimate of the strength of the temporal association between exposure and outcome. Having a complete understanding of operating

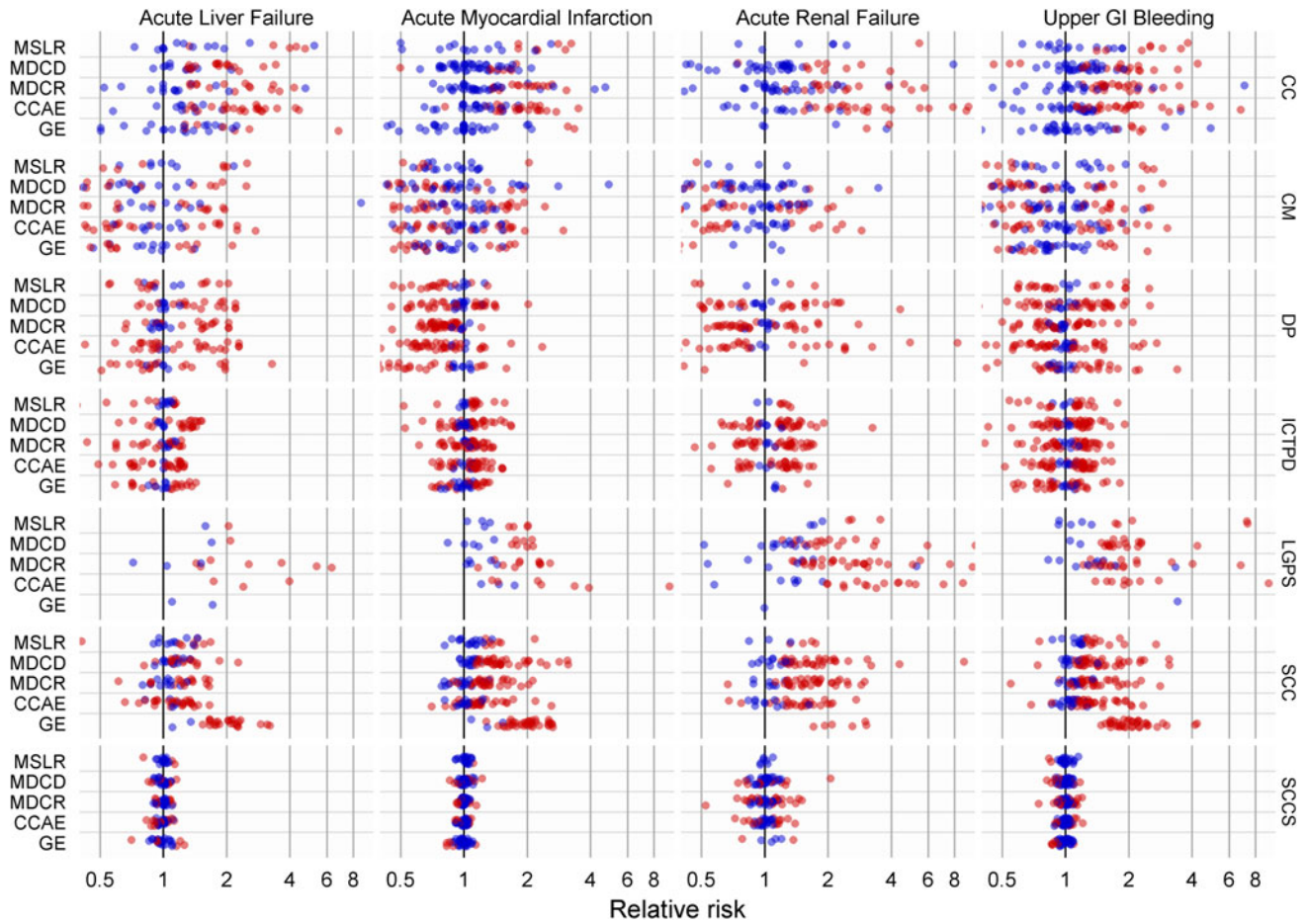

Fig. 2 Point estimates for the negative controls across all analysis methods. *Red dots* indicate estimates where the corresponding 95 % confidence interval does not include one while *blue dots* indicate estimates where the corresponding 95 % confidence interval does include zero (i.e., 95 % of all dots should be blue). *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

characteristics—including predictive accuracy (AUC), bias, mean squared error, and coverage probability—provides context for properly interpreting the results of each effect estimate. Comparing operating characteristics across alternative methods can provide evidence to support decision making about which analyses should be performed against which databases for which outcomes, which directly inform design decisions in an attempt to have resulting output being the as reliable as possible.

We believe that performance measures such as AUC should always be reported in an outcome-specific manner. Outcomes may be expected to affect estimates of varying magnitude, based on myriad factors including disease prevalence, degree of confounding, and pathophysiology of events. For example, acute myocardial infarction has a high background rate, so one would not expect many drugs to have large effects (RR $>$ 2) and smaller effect sizes with RR $<$ 1.5 may provide sufficient evidence to justify product withdrawal (e.g., rofecoxib and rosiglitazone). It is worth noting that the key evidence supporting these withdrawals was generated from randomized clinical trials and meta-analysis of study results [21–23], though several groups have tried to illustrate retrospectively that these effects could have been detected using observational healthcare database [24–29]. By contrast, acute liver injury represents a rare event less likely to be detected in clinical trials, so observing a RR $>$ 2 in the post-approval setting is less surprising, and although such a finding may trigger further evaluation, relatively large effects may still yield a tolerable benefit-risk profile (e.g., isoniazid).

This study builds on past work which has attempted to evaluate the performance and compare results across different analysis approaches. Ryan et al. [30] describes an earlier version of the OMOP experiment which evaluated the performance of different methods across ten observational databases using only 53 test cases. The magnitude of the performance in predictive accuracy was comparable to that observed in this study, though there was greater

uncertainly in our performance measures due to the limited sample of test cases. By expanding the reference set to a larger number of drugs per outcome, we were able to gain greater precision in our performance metrics and also allow us to stratify our evaluation by outcome to observe differential operating characteristics in different settings. In another study [31], data from 7 databases across 4 countries comprising over 20 million subjects in Europe were used in a distributed architecture to evaluate the performance of a wide array of methods. A reference set of 44 positive and 50 negative controls was constructed, and the authors focused exclusively on discrimination between positive and negative controls, showing good predictive accuracy across multiple analytical methods. Further work evaluating methods within this same European data network when applied to the OMOP reference set had further reinforced the findings [32]. Since multiple experiments have now successfully replicated measures of discrimination in different data environments and with different reference sets, we can begin to gain confidence in establishing benchmark expectations for the performance of traditional analysis methods. This work complements the prior study by Glanz et al. [33] which demonstrated the potential utility of many of the same designs for vaccine surveillance, by validating their performance in simulated data. Other cross-method comparisons have provided important insights about study design characteristics that may, in part, explain the observed operating characteristics reported within this study [34]. These results also help situate findings from specific drug-outcome associations where different designs have been applied to the same data to evaluate a potential effect [35–37].

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 expected to be better at addressing specific questions. For example, if the outcome of interest is most commonly observed in a hospital setting, databases with in-inpatient data may be more reliable. Each database exhibits unique limitations that could impact performance. For example, payer-based claims data may provide shorter longitudinal capture due to high turnover, while outpatient EHR systems may have more incomplete capture during the observation period.

4.1 Recommendations

How should these results influence current practice? We see value in applying an empirical approach not just to methods evaluation but to the prospective evaluation of safety concerns. The choice of the appropriate analysis should be informed not only by expert opinion but also based on evidence that the chosen analysis would perform best for the task at hand. Our recommendation for performing observational studies to identify and analyze risk is an empirical approach: for each database and outcome, perform a full evaluation of all methods and possible analysis choices as presented here, and select the analysis based on the observed operating characteristics including predictive accuracy, bias, MSE and coverage probability. These operating characteristics should also be reported when communicating the results of a study.

It is not sufficient to interpret an effect estimate in isolation. Instead, we recommend that every observational study include a set of negative control drugs and if possible positive control drugs. As our results show, often high relative risks estimates are observed even when no effect is present, and a set of control drugs can help quantify the likelihood that an observed elevated relative risk reflects a true effect. Similarly, by computing the MSE for the method used we can better appreciate how far the true effect size could be from the observed one.

4.2 Limitations

A prerequisite to supporting this type of empirical evaluation is establishing a sufficient sample of test cases with adequate confidence in the classification of positive and negative controls. In an ideal situation, the test cases would reflect the scenarios envisioned for a risk identification system, mirroring the types of confounding and other sources of bias that exist in observational healthcare data which would bedevil an epidemiologic study of a yet-unestablished drug effect. In this experiment, test cases were selected on the basis of product labeling and literature, but neither source can be considered definitive and evidence continues to evolve as real-world product exposure accumulates. If misclassification of test cases (drug-outcome pairs classified as positives which are truly negative, or negative controls which have a true effect) were present, it could result in underestimation of absolute measure of predictive accuracy but should not substantially bias the relative comparisons of methods. Our assessment of bias, mean squared error, and coverage probability is beholden to the assumption that all negative controls have a true relative risk = 1 across all exposed patients, though it is entirely conceivable that these drug-outcome pairs may have small effects which have not yet observed in past research or which may only manifest in specific patient subpopulations. Note that our analytic methods do not have access to the "labels" for the test cases, that is, whether or not 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.

A specific challenge we face in comparing methods performance is the multiplicity of analysis choices that are embedded within each method, which makes it difficult to represent a relative assessment that holds all experiment design factors constant. Whereas some analysis choices can be considered to span across multiple methods (ex. time-at-risk), many analysis choices are method-specific (ex. the new user cohort design is the only method we studied which required selection of an external comparator group). Whereas in theory it can be argued that a case–control design nested within a cohort should yield the same effect estimate as a cohort design [38], in practice, we see that the implementation of different methods encompasses a wide array of decisions which can result in different findings. In this regard, we think it is more appropriate to focus an empirical evaluation on the comparison of a fully-specified analysis with clear definition and full transparency into all analysis choices imbedded within a method, rather than to use the assessment to add to conceptual discussion about basic study designs in the abstract [39]. We believe the current comparison of methods as executed is valid, but it does raise the question of whether future method development could incorporate lessons from analysis choices in other methods to improve their performance. One advantage of applying a consistent empirical evaluation framework is that it can provide a common benchmark for current performance and a target for future methods innovation.

The generalizability of our findings remains unclear. The fact that we see different analyses yield the highest predictive accuracy for different database-outcome pairs, suggests caution is needed when projecting these results to other databases or other outcomes. Further experiments are needed to determine the degree to which results can be generalized across outcomes and databases. A possible direction is to conduct similar experiment for additional 19 outcomes identified by the Exploring and Understanding Adverse Drug Reactions (EU-ADR) project (http://www.euadr-project.org) as high-priority safety issues [40].

4.3 Future Work

A key future research question is whether or not expert analysts can outperform experimentally driven customization. This is an open question as essentially no data are currently available to shed light on the performance of expert analysts. The experiments reported in this paper do not extend customization to the specific drug and thus stop one step short of the sort of customization available to an expert analyst. Ongoing research is addressing this question although we do note that published observational studies not uncommonly apply an identical epidemiological design to multiple drugs for the same outcome.

5 Conclusions

Observational healthcare data can provide useful information for risk identification of medical product effects on acute liver injury, acute myocardial infarction, acute renal failure and gastrointestinal bleeding if appropriate methods are applied to sufficient data. Self-controlled methods demonstrated strong predictive accuracy across all databases and outcomes under study. All methods were observed to have substantial error and poor coverage, suggesting that effect estimates from all methods require calibration to ensure proper interpretation of study results. Further empirical evaluation is required to gauge the generalizability of these findings to other databases and outcomes.

Acknowledgments The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health (FNIH) through generous contributions from the following: Abbott, Amgen Inc., AstraZeneca, Bayer Healthcare Pharmaceuticals, Inc., Biogen Idec, Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Janssen Research and Development, Lundbeck, Inc., Merck & Co., Inc., Novartis Pharmaceuticals Corporation, Pfizer Inc, Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi-aventis, Schering-Plough Corporation, and Takeda. Drs. Ryan, Stang and Schuemie are employees of Janssen Research and Development. Dr. Schuemie received a fellowship from the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration. Drs. Schuemie, Suchard, Madigan, and Hartzema have received a grant previously from FNIH. Dr. DuMouchel is an employee of Oracle Health Sciences. Dr. Overhage is an employee of Siemens. Christian Reich is an employee of AstraZeneca.

This article was published in a supplement sponsored by the Foundation for the National Institutes of Health (FNIH). The supplement was guest edited by Stephen J.W. Evans. It was peer reviewed by Olaf H. Klungel who received a small honorarium to cover out-of-pocket expenses. S.J.W.E has received travel funding from the FNIH to travel to the OMOP symposium and received a fee from FNIH for the review of a protocol for OMOP. O.H.K has received funding for the IMI–PROTECT project from the Innovative Medicines Initiative Joint Undertaking (http://www.imi.europa.eu) under Grant Agreement no 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and EFPIA companies' in kind contribution.

References

1. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.
2. Berlin JA, Glasser SC, Ellenberg SS. Adverse event detection in drug development: recommendations and obligations beyond phase 3. Am J Public Health. 2008;98(8):1366–71.

3. Waller PC, Evans SJ. A model for the future conduct of pharmacovigilance. Pharmacoepidemiol Drug Saf. 2003;12(1):17–29.
4. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
5. Food and Drug Administration Amendments Act of 2007. p. Public Law 110-85, 21 STAT. 823 (2007).
6. FDA. The Sentinel Initiative: A National Strategy for Monitoring Medical Product Safety. May 2008 [cited 2012 September 15]. http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm089474.htm.
7. FDA Drug Safety Communication: Update on the risk for serious bleeding events with the anticoagulant Pradaxa (dabigatran). November 2, 2012 [cited 2012 December 1]. http://www.fda.gov/Drugs/DrugSafety/ucm326580.htm.
8. DuMouchel B, Ryan PB, Schuemie MJ, Madigan D. Evaluation of disproportionality safety signaling applied to health care databases. Drug Saf (in this supplement issue).doi:10.1007/s40264-013-0106-y.
9. Madigan D, Schuemie MJ, Ryan PB. Empirical performance of the case–control method: lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0105-z.
10. Norén GN, Bergvall T, Ryan PB, Juhlin K, Schuemie MJ, Madigna D. Empirical performance of the calibrated self-controlled cohort analysis within Temporal Pattern Discovery: lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0095-x.
11. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0099-6.
12. Ryan PB, Schuemie MJ, Madigan D. Empirical performance of a self-controlled cohort method: lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0101-3.
13. Schuemie MJ, Madigan D, Ryan PB. Empirical performance of Longitudinal Gamma Poisson Shrinker (LGPS) and Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs (LEOPARD): lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0107-x.
14. Suchard MA, Zorych I, Simpson SE, Schuemie MJ, Ryan PB, Madigan D. Empirical performance of the self-controlled case series design: lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue).doi:10.1007/s40264-013-0100-4.
15. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med. 2013. doi:10.1002/sim.5925.
16. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0097-8.
17. Greenland S. Confounding and exposure trends in case-crossover and case-time-control designs. Epidemiology. 1996;7(3):231–9.
18. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol. 1991;133(2):144–53.
19. Maclure M, Fireman B, Nelson JC, Hua W, Shoaibi A, Paredes A, et al. When should case-only designs be used for safety monitoring of medical products? Pharmacoepidemiol Drug Saf. 2012;21 Suppl 1:50–61.
20. Gagne JJ, Fireman B, Ryan PB, Maclure M, Gerhard T, Toh S, et al. Design considerations in an active medical product safety monitoring system. Pharmacoepidemiol Drug Saf. 2012;21(Suppl 1):32–40.
21. Bombardier C, Laine L, Reicin A, Shapiro D, Burgos-Vargas R, Davis B, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. N Engl J Med. 2000;343(21):1520–8 (2 p following 8).
22. Bresalier RS, Sandler RS, Quan H, Bolognese JA, Oxenius B, Horgan K, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med. 2005;352(11):1092–102.
23. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med. 2007;356(24):2457–71.
24. Cunnington M, Webb D, Qizilbash N, Blum D, Mander A, Funk MJ, et al. Risk of ischaemic cardiovascular events from selective cyclooxygenase-2 inhibitors in osteoarthritis. Pharmacoepidemiol Drug Saf. 2008;17(6):601–8.
25. Graham DJ, Campen D, Hui R, Spence M, Cheetham C, Levy G, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case–control study. Lancet. 2005;365(9458):475–81.
26. Brown JS, Kulldorff M, Chan KA, Davis RL, Graham D, Pettus PT, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. Pharmacoepidemiol Drug Saf. 2007;16(12):1275–84.
27. Brownstein JS, Sordo M, Kohane IS, Mandl KD. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. PLoS ONE. 2007;2(9):e840.
28. Brownstein JS, Murphy SN, Goldfine AB, Grant RW, Sordo M, Gainer V, et al. Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. Diabetes Care. 2010;33(3):526–31.
29. Graham DJ, Ouellet-Hellstrom R, MaCurdy TE, Ali F, Sholley C, Worrall C, et al. Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone. JAMA. 2010;304(4):411–8.
30. Ryan PB, Madigan D, Stang PE, Marc Overhage J, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. Stat Med. 2012;31(30):4401–15.
31. Schuemie MJ, Coloma PM, Straatman H, Herings RM, Trifirò G, Matthews JN, et al. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. Med Care. 2012.
32. Coloma PM, Avillach P, Salvo F, Schuemie MJ, Ferrajolo C, Pariente A, et al. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. Drug Saf. 2013;36(1):13–23.
33. Glanz JM, McClure DL, Xu S, Hambidge SJ, Lee M, Kolczak MS, et al. Four different study designs to evaluate vaccine safety were equally validated with contrasting limitations. J Clin Epidemiol. 2006;59(8):808–18.
34. Nicholas JM, Grieve AP, Gulliford MC. Within-person study designs had lower precision and greater susceptibility to bias because of trends in exposure than cohort and nested case–control designs. J Clin Epidemiol. 2012;65(4):384–93.
35. Ramsay EN, Pratt NL, Ryan P, Roughead EE. Proton pump inhibitors and the risk of pneumonia: a comparison of cohort and self-controlled case series designs. BMC Med Res Methodol. 2013;13(1):82.
36. Hubbard R, Farrington P, Smith C, Smeeth L, Tattersfield A. Exposure to tricyclic and selective serotonin reuptake inhibitor antidepressants and the risk of hip fracture. Am J Epidemiol. 2003;158(1):77–84.

37. Tata LJ, Fortun PJ, Hubbard RB, Smeeth L, Hawkey CJ, Smith CJ, et al. Does concurrent prescription of selective serotonin reuptake inhibitors and non-steroidal anti-inflammatory drugs substantially increase the risk of upper gastrointestinal bleeding? Aliment Pharmacol Ther. 2005;22(3):175–81.
38. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010;19(8):858–68.
39. Pearce N. Classification of epidemiological study designs. Int J Epidemiol. 2012;41(2):393–7.
40. Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salame G, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiol Drug Saf. 2009;18(12):1176–84.
41. Tisdale J, Miller D. Drug-induced diseases: prevention, detection, and management. 2nd ed. American Society of Health-System Pharmacists; 2010.