Drug Saf (2013) 36 (Suppl 1):S73–S82
DOI 10.1007/s40264-013-0105-z

ORIGINAL RESEARCH ARTICLE

Empirical Performance of the Case–Control Method: Lessons for Developing a Risk Identification and Analysis System

David Madigan · Martijn J. Schuemie · Patrick B. Ryan

Abstract

Background Considerable attention now focuses on the use of large-scale observational healthcare data for understanding drug safety. In this context, analysts utilize a variety of statistical and epidemiological approaches such as case–control, cohort, and self-controlled methods. The operating characteristics of these methods are poorly understood.

Objective Establish the operating characteristics of the case–control method for large scale observational analysis in drug safety.

Research Design We empirically evaluated the case–control approach in 5 real observational healthcare databases and 6 simulated datasets. We retrospectively studied the predictive accuracy of the method when applied to a collection of 165 positive controls and 234 negative controls across 4 outcomes: acute liver injury, acute myocardial infarction, acute kidney injury, and upper gastrointestinal bleeding.

Results In our experiment, the case–control method provided weak discrimination between positive and negative controls. Furthermore, the method yielded positively biased estimates and confidence intervals that had poor coverage properties.

Conclusions For the four outcomes we examined, the case–control method may not be the method of choice for estimating potentially harmful effects of drugs.

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.

D. Madigan (✉)
Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA
e-mail: david.madigan@columbia.edu

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

P. B. Ryan
Janssen Research and Development LLC, Titusville, NJ, USA

D. Madigan · M. J. Schuemie · P. B. Ryan
Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, MD, USA

1 Background

In 2007, the United States Congress passed the Food and Drug Administration (FDA) Amendment Act, [1] which called for the establishment of an "active postmarket risk identification and analysis system" with access to patient-level observational data from 100 million lives by 2012 [2]. It is envisioned that such a system would "use sophisticated statistical methods to actively search for patterns in prescription, outpatient, and inpatient data systems that might suggest the occurrence of an adverse event, or safety signal, related to drug therapy [3]." Observational healthcare data, such as administrative claims and electronic health records, have previously been used to study specific effects of a particular medical product and subsequent health outcomes of interest. Recent advances in health informatics and large-scale analytics have raised the potential to expand the use of these same data beyond the currently reactive and customized one-off investigations of specific questions to provide a proactive and systematic process for monitoring all regulated products for a wide array of health outcomes of interest. To achieve this objective, several outstanding questions require resolution before a risk identification and analysis system can be properly implemented and embedded within safety decision-making processes, including 'which analytical methods to apply?', 'which data should be used?', and 'how credible is the evidence from observational analyses?'

Several methods have been proposed for use in a risk identification system, but little empirical research exists to inform the expected operating characteristics of these approaches. For example, how successful are different methods at discriminating between causal and non-causal associations? What is the coverage of 95 % confidence intervals produced by the methods? What level of bias is typically observed? One such method is the classical case–control method and the recent medical literature features many case–control analyses of drug safety issues using large-scale observational data. We tested the case–control approach in 5 real observational healthcare databases and 6 simulated datasets, retrospectively studying the predictive accuracy of the method when applied to a collection of 165 positive controls and 234 negative controls across 4 outcomes: acute liver injury, acute myocardial infarction, acute kidney injury, and upper gastrointestinal bleeding. We estimate how well the methods can be expected to identify true effects and discriminate between true effects and true non-effects.

2 Methods

2.1 Overview of the Case–Control Method

The case–control method [2–4] remains one of the most widely used methods in observational studies of the effects of drugs. Case–control studies consider the question "are persons with a specific condition exposed more frequently to a specific drug than those without the disease?" Thus, the central idea is to compare "cases," i.e., individuals that experience the outcome of interest with (possibly) matched "controls," i.e., individuals that did not experience the outcome of interest. The comparison focuses on differential exposure to the drug of interest in the two groups; greater exposure amongst the cases than amongst the controls suggests a possible positive association.

The case–control method was developed in situations where data on subjects was costly to acquire, and study budgets did not allow for recruiting and following large cohorts [5]. Nowadays, case–control studies are typically nested in a cohort, such as the population in a longitudinal observational database, where there is little cost in retrieving data on more subjects. In such databases, the point in time a case experiences the outcome is referred to as the "index date," and exposure is considered on or prior to the index date. Similarly, the case–control method must select an index date for each control, and this is typically the same calendar day as the index date of the corresponding case.

Several design choices must be made when performing a case–control analysis. In order to ascertain exposure prior to the index date, the case–control analyses require a minimum period of observation (MPO) such as 180 days prior to this date. All patients in the database who did *not* experience the outcome represent potential controls. To create a set of controls for a given case set, standard practice divides all cases into sub-groups defined by "matching variables" such as sex and age. In our implementation, for each index date, we randomly select controls from the pool of potential controls with the same values for the matching variables who were observed on this index date and with an observation period that extends at least MPO days prior to the index date. Each selected control adopts this index-date as a "control index-date." Analysts must choose the number of controls to attempt to match to each case—5 or 10 controls per case is typical.

A drug contributes to the exposed count for a particular patient in the case group or control group if the index date (control index date for a control) falls within some interval of time related to drug usage. For example, an analyst might decide that a patient contributes to the exposed count only if the index date is within 30 days from the drug initiation date. Or, an analyst might consider index dates that occur anytime during the drug exposure or within some time period after the end of the drug exposure. These time windows reflect a priori beliefs about the temporal relationship between the exposure and the outcome.

Following Rothman et al. [5] denote by A_1 the number of exposed cases and by A_0 the number of unexposed cases. Similarly denote by B_1 the number of exposed controls and by B_0 the number of unexposed controls. The case–control estimate of effect size is given by the odds ratio:

$$\frac{A_1/B_1}{A_0/B_0}.$$

Let T_1 denote exposed person-time and T_0 be unexposed exposure time. If the controls are selected independently of exposure, one would expect B_1/T_1 to be approximately equal to B_0/T_0. Then:

$$\frac{A_1/B_1}{A_0/B_0} = \frac{A_1/(B_1/T_1)T_1}{A_0/(B_0/T_0)T_0} \cong \frac{A_1/T_1}{A_0/T_0},$$

so that the case–control estimate is approximately equal to the ratio of incidence rates in the source population. In practice, analysts often adjust for potential confounders by using a logistic regression to estimate the odds ratio. Note however that because subjects may have been exposed to the target drug earlier than MPO days prior to the index date, inadvertently adjusting for "intermediate variables," i.e., variables on the causal pathway between the target drug and the target outcome, can occur and thus introduce another source of bias.

2.2 Additional Features

Our publicly available implementation of the case–control method includes a number of features such as selecting controls by matching on visit dates, restricting analysis to the first occurrence of each drug, nesting within an indication, and using an option for either conditional logistic regression or Bayesian logistic regression. When logistic regression is used for analysis, a number of additional covariates may be included into the model, for example the number of drugs that the person has taken, the number of conditions experienced by the patient, the number of visits, and/or the Charlson comorbidity index, although we do not explore these possibilities in what follows [6]. See the technical specification below for the set of parameters that pertain to each available option.

2.3 Experimental Design

We used five observational healthcare databases to evaluate the performance of the case–control approach. Using multiple databases facilitates performance comparisons across different populations and data capture processes. The databases are: MarketScan® Research Databases including—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), and the General Electric Centricity™, (GE, 11.2 m persons) database. GE is an electronic health record (EHR) database while the other four databases contain administrative claims data. We also constructed a 10 million-person simulated dataset using the OSIM2 simulator [7] to model the MSLR database, and replicated this 6 times to allow for injection of signals of known size (relative risks = 1, 1.25, 1.5, 2, 4, 10). We have provided a more detailed description of the data elsewhere [8].

We executed the case–control method using all 385 parameter combinations against 399 drug–outcome pairs to generate an effect estimate and standard effort for each pair-parameter-database combination. 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. The full set of test cases and its construction is described elsewhere [9]. For every database we restricted the analysis to 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 [4].

2.4 Metrics

The estimates and associated standard errors for all of the analyses are available for download at: http://omop.org/Research. To gain insight into the ability of a method to distinguish between positive and negative controls, we computed the area under the receiver operating characteristic curve (AUC), a measure of predictive accuracy [6], for every analysis. 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 whether a method produces correct relative risk estimates, we must know the true effect size. In real data, this true effect size is never known with great accuracy for positive controls, and we must restrict our analysis to the negative controls where we assume that the true relative risk is 1. Fortunately, in the simulated data sets we do know the true relative risk for all injected signals. Using both the negative controls in real data, and injected signals in the simulated data, we compute the coverage probability: the percentage of confidence intervals that contain the true relative risk. In case of an unbiased estimator with accurate 95 % confidence interval estimation we would expect the coverage probability to be 95 %.

Lastly, we are interested in the extent to which each parameter can influence the estimated relative risk. For every parameter, we evaluated how much the estimated relative risk changed as a consequence of changing a single parameter while keeping all other parameters constant.

3 Results

3.1 Predictive Accuracy of All Settings

Figure 1 shows the predictive accuracy, as measured by AUC, of all case–control analysis choices across the four outcomes and five databases. Overall, the case–control

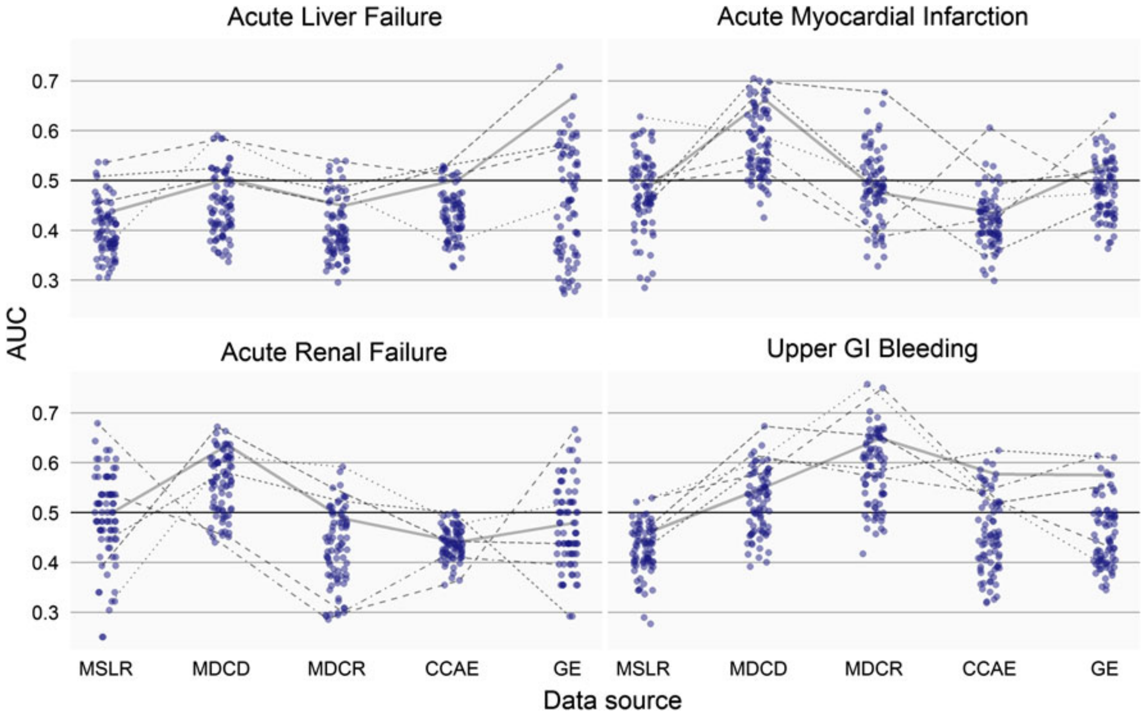


Fig. 1 Area under ROC curve (AUC) for case–control parameters, by outcome and database. *Each dot* represents one of the unique parameter combination of the case–control method. The *solid grey line* highlights the parameter that had the highest average AUC across all 20 outcome-database scenarios. The *dashed lines* identify setting with the highest AUC in at least one database within each outcome.*AUC* area under ROC curve, *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

method often has poor performance, with AUCs around or below 0.5, the equivalent of random guessing. However, for certain outcome-database combinations, there exist parameter settings for which the AUC performance approaches 0.70. These include all acute liver failure and acute renal failure in GE, AMI and Upper GI bleeding in MDCR, AMI, acute renal failure and Upper GI bleeding in MDCD, and acute renal failure in MSLR.

For each outcome-database scenario Table 1 provides the parameter settings that yielded the highest AUC. Remarkably, 17 different settings proved optimal and three scenarios proved optimal in more than one scenario. Analysis 2000225 proved optimal for both acute liver injury and acute kidney injury in MDCD, Analysis 2000321 was optimal for upper GI bleeding in MDCD and acute kidney injury in MDCR, and Analysis 2000219 was optimal for acute liver injury in MDCR and MSLR.

None of the optimal settings included the index date in the time at risk and almost all computed the target odds ratio with Mantel–Haenszel adjustment. The dashed and dotted lines in Fig. 2 indicate the performance of these top-performing setting across databases within the same outcome, showing that the optimal setting for one database can sometimes perform poorly when used for another database in the same outcome.

3.2 Overall Optimal Settings

The parameter setting with the best average performance across the 20 outcome-database scenarios is highlighted in the shaded grey line in Fig. 1, and represents analysis 2000031. The analyses settings corresponding to 2000031 use up to 100 controls per case, require a 30-day observation period prior to outcome, define the at-risk period as 30-days from exposure start, and include the index date in this time-at-risk, use age and sex to match controls, nest within indication, and compute the effect estimate using Mantel–Haenszel adjustment for age and sex. This setting performed well for some outcomes in some databases although it was never the optimal analysis for any one database-outcome combination. In the remainder of this paper we will use 2000031 as the representative setting for the case–control method.

The Appendix contains the effect estimates for all test cases across the five databases using this optimal parameter setting (2000031). To illustrate patterns in these findings, we discuss four specific test cases, as shown in Fig. 2. Naproxen is a non-steroidal anti-inflammatory drug (NSAID), and is known to be associated with upper GI bleeding [10]. An increased risk was therefore correctly identified in all five of the databases, although this effect

Table 1 Optimal analysis choices for the case–control method, by database and outcome

Data	Acute liver injury	Acute renal failure	Acute myocardial infarction	Upper GI bleeding
CCAE	AUC = 0.53 (CC: 2000027)	AUC = 0.5 (CC: 2000037)	AUC = 0.61 (CC: 2000195)	AUC = 0.62 (CC: 2000314)
	Up to 10 controls per case	Up to 10 controls per case	Up to 10 controls per case	Up to 10 controls per case
	Min. time before outcome: 30 days	Min. time before outcome: 30 days	Min. time before outcome: 30 days	Min. time before outcome: 30 days
	Time-at-risk: 30 days from exp. start	Time-at-risk: 30 days from exp. start	Time-at-risk: 30 days from exp. start	Time-at-risk: exposure + 30 days
	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no
	Match on age and sex	Match on age, sex and visit (30 days)	Match on age and sex	Match on age and sex
	Nesting within indication: yes	Nesting within indication: yes	Nesting within indication: no	Nesting within indication: yes
	First exposure only	First exposure only	First exposure only	All exposures
	OR adjusted for age and sex	OR adjusted for age and sex	OR adjusted for age and sex	Unadjusted odds ratio (OR)
GE	AUC = 0.73 (CC: 2000032)	AUC = 0.67 (CC: 2000291)	AUC = 0.63 (CC: 2000007)	AUC = 0.61 (CC: 2000313)
	Up to 100 controls per case	Up to 10 controls per case	Up to 100 controls per case	Up to 10 controls per case
	Min. time before outcome: 30 days	Min. time before outcome: 180 days	Min. time before outcome: 30 days	Min. time before outcome: 180 days
	Time-at-risk: 30 days from exp. start	Time-at-risk: exposure + 30 days	Time-at-risk: 30 days from exp. start	Time-at-risk: exposure + 30 days
	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no
	Match on age and sex	Match on age and sex	Match on age and sex	Match on age and sex
	Nesting within indication: yes	Nesting within indication: yes	Nesting within indication: no	Nesting within indication: yes
	First exposure only	First exposure only	First exposure only	All exposures
	Unadjusted odds ratio (OR)	OR adjusted for age and sex	OR adjusted for age and sex	OR adjusted for age and sex
MDCD	AUC = 0.59 (CC: 2000225)	AUC = 0.67 (CC: 2000225)	AUC = 0.70 (CC: 2000039)	AUC = 0.67 (CC: 2000321)
	Up to 10 controls per case	Up to 10 controls per case	Up to 10 controls per case	Up to 10 controls per case
	Min. time before outcome: 180 days	Min. time before outcome: 180 days	Min. time before outcome: 30 days	Min. time before outcome: 180 days
	Time-at-risk: 30 days from exp. start	Time-at-risk: 30 days from exp. start	Time-at-risk: 30 days from exp. start	Time-at-risk: exposure + 30 days
	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no
	Match on age, sex and visit (30 days)	Match on age, sex and visit (30 days)	Match on age, sex and visit (180 days)	Match on age, sex and visit (30 days)
	Nesting within indication: yes	Nesting within indication: yes	Nesting within indication: yes	Nesting within indication: yes
	All exposures	All exposures	First exposure only	All exposures
	OR adjusted for age and sex	OR adjusted for age and sex	OR adjusted for age and sex	OR adjusted for age and sex
MDCR	AUC = 0.54 (CC: 2000219)	AUC = 0.59 (CC: 2000321)	AUC = 0.68 (CC: 2000111)	AUC = 0.76 (CC: 2000326)
	Up to 10 controls per case	Up to 10 controls per case	Up to 10 controls per case	Up to 10 controls per case
	Min. time before outcome: 180 days	Min. time before outcome: 180 days	Min. time before outcome: 30 days	Min. time before outcome: 180 days
	Time-at-risk: 30 days from exp. start	Time-at-risk: exposure + 30 days	Time-at-risk: exposure + 30 days	Time-at-risk: exposure + 30 days
	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no
	Match on age and sex	Match on age, sex and visit (30 days)	Match on age, sex and visit (180 days)	Match on age, sex and visit (30 days)
	Nesting within indication: yes	Nesting within indication: yes	Nesting within indication: no	Nesting within indication: yes
	First exposure only	All exposures	First exposure only	First exposure only
	OR adjusted for age and sex	OR adjusted for age and sex	OR adjusted for age and sex	Unadjusted odds ratio (OR)

Table 1 continued

Data	Acute liver injury	Acute renal failure	Acute myocardial infarction	Upper GI bleeding
MSLR	AUC = 0.54 (CC: 2000219)	AUC = 0.68 (CC: 2000003)	AUC = 0.63 (CC: 2000316)	AUC = 0.53 (CC: 2000325)
	Up to 10 controls per case	Up to 10 controls per case	Up to 10 controls per case	Up to 10 controls per case
	Min. time before outcome: 180 days	Min. time before outcome: 30 days	Min. time before outcome: 180 days	Min. time before outcome: 180 days
	Time-at-risk: 30 days from exp. start	Time-at-risk: 30 days from exp. start	Time-at-risk: exposure + 30 days	Time-at-risk: exposure + 30 days
	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no	Index date in time-at-risk: no
	Match on age and sex	Match on age and sex	Match on age and sex	Match on age, sex and visit (30 days)
	Nesting within indication: yes	Nesting within indication: no	Nesting within indication: yes	Nesting within indication: yes
	First exposure only	First exposure only	First exposure only	First exposure only
	OR adjusted for age and sex	OR adjusted for age and sex	Unadjusted odds ratio (OR)	OR adjusted for age and sex

AUC area under ROC curve, *OR* odds ratio, *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

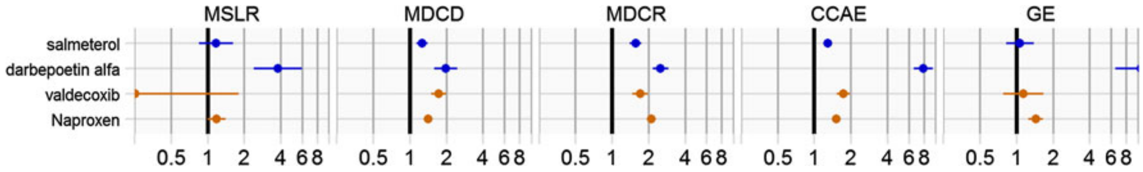


Fig. 2 Odds ratio and 95 % confidence interval for 4 example drugs and upper GI bleeding or acute renal failure across databases, using the overall optimal settings. *OR* odds ratio, *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

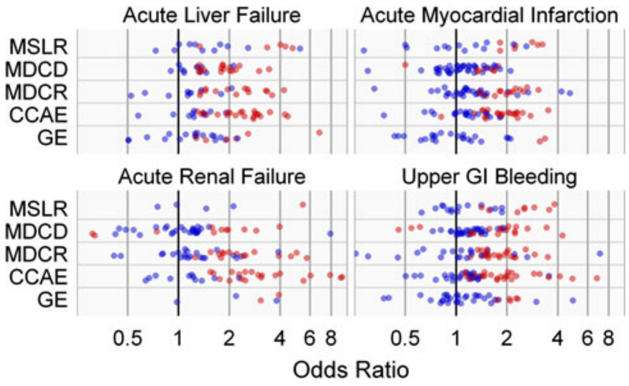


Fig. 3 Estimates for the negative control drugs, where the assumed true relative risk is one, using those settings that achieved the highest AUC averaged over all databases and outcomes. *Red* indicates relative risks that are statistically significant different from 1. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

was statistically significant in only four of the five databases. Valdecoxib is another NSAID that is assumed to have a lower risk of GI bleed than naproxen [11]. However, only three of the five databases yield an estimated effect size that is statistically significant, but not in the other two. This may be due to gastro-protective strategies such as co-prescribing with proton-pump inhibitors, although why this might happen in two of the five databases and not the other three is unclear. All five databases yield a positive estimate of the association between salmeterol and acute myocardial infarction; the estimated risk is close to two and statistically significant in three of the five databases. No previous evidence suggests a causal association in this context. Since salmeterol is indicated for chronic asthma and chronic obstructive pulmonary disease, perhaps uncontrolled protopathic bias explains the findings. Darbepoetin alfa shows a positive statistically significant relationship with acute renal failure across all five databases, despite the fact that this drug–outcome pair is also one of our negative controls. Since darbepoetin alfa is used for the treatment of anemia in patients with chronic renal failure, presumably this is another example of protopathic bias.

3.3 Bias

Figure 3 shows the magnitude of bias observed across the estimates for the negative control test cases in the five real databases. We see across all four outcomes and all 5 databases that the case–control method is positively biased, that is, the expected value for the method when applied to a negative control is greater than 1.

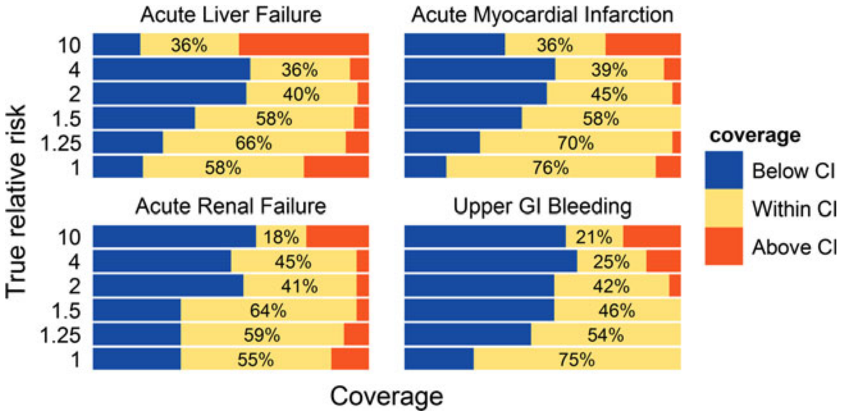


Fig. 4 Coverage probability of case–control method at different levels of true effect size, by outcome

Table 2 Sensitivity to analysis choices, showing the distribution of how much the effect estimate changes when changing a single analysis choice whilst keeping all others constant

Analysis choice	q10 delta	q50 delta	q90 delta
Nesting within indicated population?	1.05	1.32	2.31
Require control to have a visit on the index date	1.03	1.3	2.11
Data source	1.04	1.26	1.96
Risk window definition	1.02	1.17	1.76
Controls per case (10 or 100)	1.02	1.11	1.39
Required observation time prior to outcome (30 or 180 days)	1.01	1.07	1.34
Exposures to include: first or all	1.01	1.07	1.4
Correct for age and sex	1.01	1.06	1.29
Maximum number of days between visit and case index date	1	1.03	1.21

3.4 Coverage Probability

Figure 4 shows the coverage probabilities on simulated data. In general, the coverage is low, with the true effect size often falling below the estimated confidence interval. As the true effect size increased, more often the true effect size was below the confidence interval, but in no scenarios did the case–control method achieve a coverage probability greater than 76 %. Coverage probabilities generally trended lower as the true injected relative risk increased—for a relative risk of 10, the coverage probabilities for the four outcomes ranged from 18 to 36 %.

3.5 Sensitivity to Analysis Choices

Table 2 shows how sensitive effect estimates were to the analysis choices. Nesting the study in the population of patients that have the primary indication of the drug has the largest effect on the estimates when compared to using the whole population: The median change in effect estimates is 32 %. In other words, when holding all other analysis choices constant, there is a 50 % chance that the observed odds ratio will change at least 32 % either positively or negatively when switching from the general population to that part of the population that has the indication. Perhaps this is to be expected given that nesting within indication is one of the primary devices analysts use to address confounding by indication. There is a 10 % chance that the impact of nesting on the odds ratio will be 131 % or more. The estimates were less sensitive to other analysis choices.

4 Discussion

This paper presents a large-scale empirical evaluation of the case–control method for estimating the effects of drugs on large observational databases. Our findings indicate that in an absolute sense the case–control method generally performs poorly in terms of its ability to discriminate between positive and negative controls. Our work also provides insight into the impact of different analysis choices within the case–control method.

The level of heterogeneity due to analysis choices focuses attention on the need to make such choices carefully. The diversity of the optimal choices suggests that the task is not simple, presumably depending on subtle issues related to biological mechanism, selection bias, and data quality to name a few. Future methods research needs to focus on approaches that are less sensitive (or what Tukey called “resistant”) to analyst choices [12]. We note that some authors have pointed out that case–control and cohort analyses ought to arrive at similar conclusions [13, 14]. While they may be correct conceptually, due to issues of covariate construction, time windows, matching strategy, etc., in practice, the designs can yield very different effect estimates and warrant evaluation separately.

Our work has a number of limitations. The value of our experimental findings depends on the correctness of the “positive” and “negative” labels we assigned to the control drug–outcome pairs [9]. The range of analytic choices we included in our experiment represent typical choices seen in the literature but certainly do not include all possible choices. While our experiments did include four administrative claims databases, we only included one EHR database. Experimental results may be different on other databases. Similarly we only included four outcomes; performance results on other outcomes may be different.

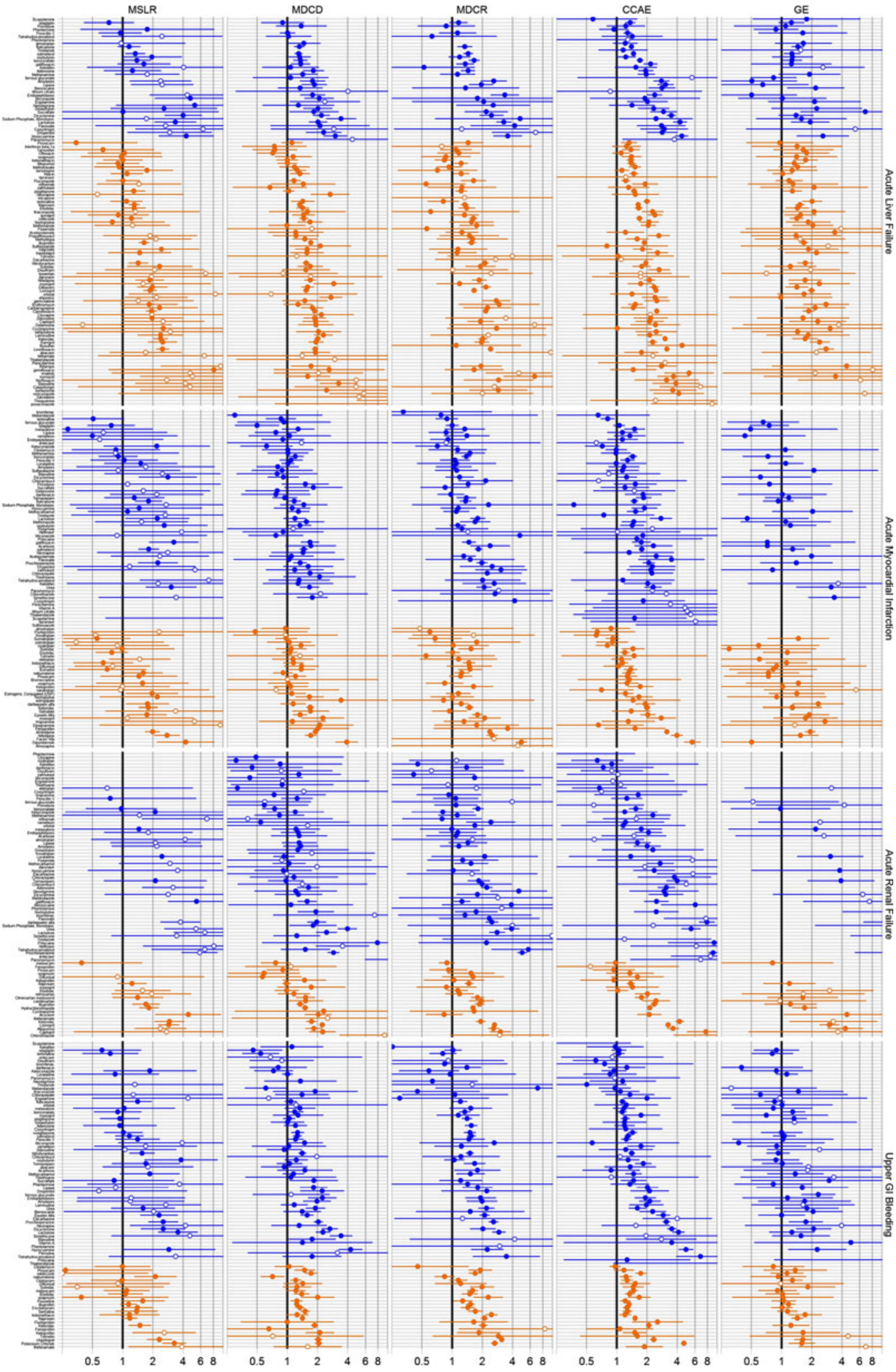
5 Conclusion

The performance of the case control method is sensitive to specific implementation choices with some settings performing quite well for one outcome in one database but the same settings performing poorly for the same outcome in a different database or for a different outcome in the same database. Secondly, the case–control method exhibits positive bias across all outcomes and all databases. For negative controls specifically, exposure is more common amongst cases than controls, indicating the inability of the matching criteria to control for morbidity and possibly other exposure-related factors. Since we utilized the kinds of matching criteria commonly seen in the literature, presumably this positive bias bedevils many published case–control studies. Finally, standard case–control confidence intervals exhibit poor coverage. We believe this is the first evaluation of its kind and our hope is that empirical evaluation becomes a routine practice in observational epidemiology.

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 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 and Madigan have both received a grant previously from FNIH.

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 (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.

Appendix



The effect estimates for all test cases across the five databases using the optimal analysis choice setting (2000031). *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

References

1. Food and Drug Administration Amendments Act of 2007. Pub. L. 110-85, 21 U.S., 2007.
2. Woodward M. Epidemiology study design and data analysis. London: Chapman & Hall/CRC; 1999.
3. Agresti A. Categorical data analysis. Hoboken: Wiley-Interscience; 2002.
4. Breslow NE, Day NE. Statistical methods in cancer research. In: The analysis of case–control studies, vol. I. France: International Agency for Research on Cancer; 1993.
5. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkens; 2008.
6. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373–83.
7. Ryan PB, Schuemie M. Evaluating performance of risk identification methods through a large-scale simulation of observational data. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0110-2
8. Overhage JM, Ryan PB, Schuemie MJ, Stang PE. Desideratum for evidence based epidemiology. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0102-2
9. 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
10. Wolfe MM, Lichtenstein DR, Singh G. Gastrointestinal toxicity of nonsteroidal antiinflammatory drugs. N Engl J Med. 1999;340(24):1888–99.
11. Makarowski W, Zhao WW, Bevirt T, Recker DP. Efficacy and safety of the COX-2 specific inhibitor valdecoxib in the management of osteoarthritis of the hip: a randomized, double-blind, placebo-controlled comparison with naproxen. Osteoarthr cartil OARS Osteoarthr Res Soc. 2002;10(4):290–6.
12. Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol. 2002;31(2):422–9.
13. Hofler M. Causal inference based on counterfactuals. BMC Med Res Methodol. 2005;5:28.
14. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010;19(8):858–68.