Drug Saf (2013) 36 (Suppl 1):S171–S180
DOI 10.1007/s40264-013-0110-2

ORIGINAL RESEARCH ARTICLE

Evaluating Performance of Risk Identification Methods Through a Large-Scale Simulation of Observational Data

Patrick B. Ryan · Martijn J. Schuemie

Abstract

Background There has been only limited evaluation of statistical methods for identifying safety risks of drug exposure in observational healthcare data. Simulations can support empirical evaluation, but have not been shown to adequately model the real-world phenomena that challenge observational analyses.

Objectives To design and evaluate a probabilistic framework (OSIM2) for generating simulated observational healthcare data, and to use this data for evaluating the performance of methods in identifying associations between drug exposure and health outcomes of interest.

Research Design Seven observational designs, including case–control, cohort, self-controlled case series, and self-controlled cohort design were applied to 399 drug-outcome scenarios in 6 simulated datasets with no effect and injected relative risks of 1.25, 1.5, 2, 4, and 10, respectively.

Subjects Longitudinal data for 10 million simulated patients were generated using a model derived from an administrative claims database, with associated demographics, periods of drug exposure derived from pharmacy dispensings, and medical conditions derived from diagnoses on medical claims.

Measures Simulation validation was performed through descriptive comparison with real source data. Method performance was evaluated using Area Under ROC Curve (AUC), bias, and mean squared error.

Results OSIM2 replicates prevalence and types of confounding observed in real claims data. When simulated data are injected with relative risks (RR) ≥ 2, all designs have good predictive accuracy (AUC > 0.90), but when RR < 2, no methods achieve 100 % predictions. Each method exhibits a different bias profile, which changes with the effect size.

Conclusions OSIM2 can support methodological research. Results from simulation suggest method operating characteristics are far from nominal properties.

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

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

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

1 Background

The secondary use of observational healthcare data, such as administrative claims and electronic health records, has the potential to support the characterization of causal associations between medical product exposures and subsequent health outcomes of interest. These data are often used in

pharmacoepidemiology studies to estimate the strength of association as an average treatment effect. There has been only limited evaluation of statistical methods for identifying safety risks of drug exposure in observational healthcare data [1, 2]. Methods validation can be achieved through empirical evaluation by executing methods against observational databases in a variety of different scenarios and comparing the methods estimates to some referent standard of ground truth. Various operating characteristics can be measured as part of a methods validation exercise, including predictive accuracy, bias, and mean squared error. A key challenge in conducting methodological research in real data is establishing the ground truth effect size that can serve as the basis for comparison. Even establishing a ground truth set with dichotomous indicators for positive association or not represents a substantial challenge, requires expert consensus and subjective clinical judgment of limited available evidence from disparate sources, can result in misclassification, and is insufficient for estimating measurement error.

Simulation can complement empirical evaluation in real data by providing a known referent standard, by creating datasets with a user-defined effect size that can be modeled in the data and used to compare with method estimates. The primary limitation of simulations is that they represent simplified models of the real-world phenomenon that challenge methods, such as complex patterns of comorbidities, concomitant drug use, confounding by indication, and irregularly spaced temporal data. In this study, we describe the design and validation of a simulation of observational databases, and then use this simulation as a tool to evaluate the performance of different analysis methods in their ability to establish the strength of association between drug exposure and select health outcomes of interest.

2 Methods

2.1 Simulation Design

A more detailed description of OSIM2 is published elsewhere [3]. The full specifications, source code, and output of the OSIM2 are publicly available online at: http://omop.org/OSIM2.

Traditionally, simulation studies require fully specified models of all variables and their dependencies [4, 5]. Instead, OSIM2 derives the list of variables (i.e. patient demographics, drugs, and diagnoses), their frequencies and marginal dependency statistics from a real database. The resulting model can then be used to generate datasets of varying sizes in four consecutive steps. The generic OSIM2 model is shown in Fig. 1. In this experiment, we construct

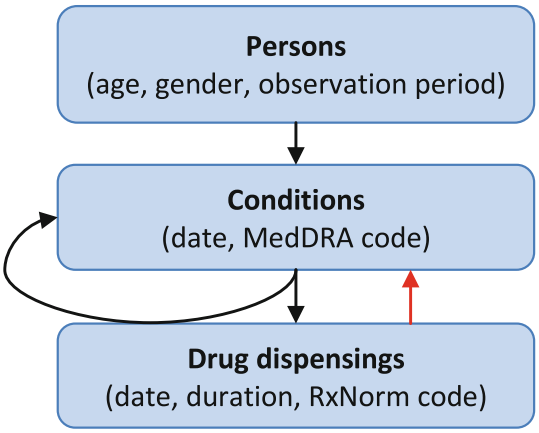

Fig. 1 Generic OSIM2 model. *Black arrows* indicate marginal dependencies derived from real data. The *red arrow* from drug dispensings to conditions represents the injected adverse drug reactions, which have to be specified by the user

a 10-million person simulated dataset from information extracted from the MarketScan® Lab Supplemental database (MSLR). MSLR contains 1.2 million patients, representing a privately insured population that has at least one recorded laboratory value, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results.

As a first step, a simulated population of a user-defined size is constructed using the age and gender distributions from the source data. Each simulated person is then modeled to have one observation period—representing the span of time for which conditions and drugs could be recorded in the database—based on the empirical distribution of observation period lengths observed from the source data. Each person is assigned a distinct condition count, as a general measure of wellness of each person. On the basis of a person's age, gender, length of observation, and condition count, the simulation then assigns specific conditions and distributes the diagnosis records throughout the observation period.

The second stage of modeling conditions follows a first-order Markov process, where the transition probabilities are empirically derived as the probability of having the next disease amongst all possible alternative diseases (as represented by 3,904 Medical Dictionary for Regulatory Activities, MedDRA® preferred terms), conditional on the prior disease and patient demographics. So, for example, from the real data, we may observe that 50 year-old males with 5 distinct conditions have a high likelihood of having their first disease in the record be 'diabetes' and a low likelihood of their first disease being 'breast cancer'. To populate the simulated 50-year-old male's condition record, we sample from the empirical distribution to select the first condition, and have a higher likelihood of assigning the person a 'diabetes' record. To determine the

second disease in that simulated person's record, we observe from the real data that 50-year-old males with 5 distinct conditions and a prior diagnosis of 'diabetes' are likely to have subsequent 'hypertension', and as a result, 'hypertension' has a high point probability in the empirical distribution. Assuming 'hypertension' had been randomly assigned as the second condition, the Markov model follows the transition probability to select a third condition for the person based on the empirical distribution for 50-year-old males with 5 distinct conditions and a prior diagnosis of 'hypertension'. This process continues until all distinct conditions are selected and distributed across the observation period.

As a third stage in the modeling process, the simulated patients are assigned drug exposures following a Markov model defined by age, gender, and prior conditions. Continuing the example, we determine the probability of observing each unique drug, as represented by 1,338 RxNorm ingredient concepts, following diagnosis of 'diabetes' among 50-year-old males. We sample from these empirical distribution, and have a high likelihood of selecting some hyperglycemic treatment, such as 'metformin'. If selected, one or more exposures of 'metformin' are then added to the person record some time following the 'diabetes' diagnosis, and is assigned exposure lengths based on the distribution of days supply of metformin from the source data. The model then continues to determine if a drug is assigned following the diagnosis of 'hypertension' and each successive condition that was previously modeled. This modeling approach guarantees that drug exposures are independent of future outcomes, conditional on past diseases. This conditioning allows the simulated data to demonstrate much of the bias and confounding we expect to observe in real data, but preserves the ability for users to prospectively define the effects of exposures on outcomes in the fourth stage—signal injection.

In the fourth stage, additional conditions are injected based on previous drug exposures. These dependencies represent adverse drug reactions, and are not derived from the real data but have to be specified by the user. OSIM2 allows users to define drug effects by the magnitude of the effect size (as a relative risk, RR) and the type of temporal relationship between drug and outcome. Prior to injection, all drug-outcome pairs have a true RR = 1, because the modeling of a drug exposure had no impact on subsequent conditions (since all baseline conditions are modeled prior to assigning drug exposures). We rely on this modeling structure to ensure that we can measure performance using the same negative controls that were used in our real data experiments. All negative control drug-outcome pairs are present in the simulation, and assumed to have RR = 1, conditional on prior conditions and drugs. For this simulation experiment, we created 5 instances of simulated datasets with injected signals for each of the positive controls in the OMOP experiment, described elsewhere [6]. Each instance defined the type of effect as insidious, meaning the drug had a constant effect on the outcome during the entire duration of drug exposure. We varied the effect size as RR = 1.25, 1.5, 2, 4, and 10. To illustrate how signals are injected, consider the positive control of 'naproxen' and 'gastrointestinal bleeding (UGIB)'. To inject an insidious signal of RR = 4, the model first identifies the number of persons with an exposure to naproxen, the proportion of those persons who already had a background condition of UGIB in their record during the simulated naproxen exposure period. We find that 3,887 of 663,402 simulated naproxen users (0.59 %) in our 10 million person dataset had a simulated diagnosis of UGIB during their exposure. In a simulated dataset prior to injecting signals, this background rate represents the counterfactual incidence of events, if naproxen had no effect on the onset of UGIB. The RR = 4 is injected by finding the number of additional people who will have the outcome attributable due to the drug; in this case, (RR − 1) * background rate * exposed population = (4 − 1) * 0.59 % * 663,402 = 11,678 persons, and randomly assign events during the period of exposure among those exposed persons who did not already have the background condition. We repeat this process for each of the positive control drugs for four outcomes: upper gastrointestinal bleeding (UGIB), acute myocardial infarction (AMI), acute liver injury (ALI), and acute renal failure (ARF). As with real observational databases, the diagnosis records due to injection are not distinguishable from the records containing background conditions, and the task of each analytical method in our experiment is to recover the true effect size from the entire dataset.

2.2 Simulation Validation

Murray et al. [3] described prior validation efforts of the simulation model. Here we compute descriptive statistics about the population characteristics in MSLR and OSIM2 to evaluate consistency. These characteristics include: age and gender distribution, average length of observation, number of drugs records and unique drugs per person, number of condition records and unique conditions per person, and the prevalence of each drug and condition. To illustrate how OSIM models some of the confounding that exists in MSLR, we provide three motivating examples of drug-outcome associations with known characteristics and compare effect estimates in both OSIM and MSLR. For each example, we estimate incidence rate ratio and 95 % confidence intervals by comparing the rate of events during the exposed time with the rate of events among the unexposed population, and stratified estimates by age and indication.

2.3 Methods Performance Experiment

Six simulated databases were created to be used to evaluate methods performance in estimating the strength of association between drug exposure and outcome. Seven methods (i.e., overall study designs) were evaluated and a detailed description and analysis of each method is provided in this supplement.

- **Case–control (CC)** compares the rate of exposure prior to outcomes with the rate of exposure in patients without outcomes [7].
- **Cohort method (CM)** is a new-user cohort design. New users of the target drug are identified using a predefined minimum period of non-use, and are compared to new users of a comparator drug or group of drugs relative risk can be adjusted for baseline covariates through various strategies, including propensity score matching [8].
- **Disproportionality methods (DP)** is a suite of methods borrowed from data-mining in spontaneous reports, including PRR (proportional reporting ratio), ROR (reporting odds ratio), BCPNN (Bayesian Confidence Propagation Neural Networks) and MGPS (Multi-item Gamma Poisson Shrinker) [9].
- **Information Component Temporal Pattern Discovery (ICTPD)** compares the disproportionality of events during a post-exposure period with the disproportionality of events during one or more pre-exposure periods to produce a self-controlled-adjusted measure of temporal association [10].
- **Longitudinal Gamma Poisson Shrinker (LGPS)** compares the incidence rate during exposure to the drug of interest to the background incidence rate, optionally applying Bayesian shrinkage. This method is often combined with **Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs (LEOPARD)**, a technique for detecting and discarding spurious signal caused by protopathic bias [11].
- **Self-Controlled Cohort (SCC)** implemented in the Observational Screening (OS) package estimates the strength of association by comparing the post-exposure incidence rate with the pre-exposure incidence rate among the patients exposed to the target drug of interest [12].
- **Self-controlled case series (SCCS)** focuses on time exposed/unexposed to target drug and occurrences of target condition. It is basically a Poisson regression condition on the person [13].

Within each method, there are multiple analysis choices that need to be made to fully specify an analysis, such as the definition of the risk window, selection of variables to use in multivariate adjustment, or strategy for selecting a comparator drug, and we have evaluated each method using multiple alternative choices of these parameters. Each combination of analysis choices was executed against 399 drug-outcome pairs to generate an effect estimate and standard error for each pair, method, and parameter 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. By construction in the simulation, no drugs cause any outcomes unless injected as a signal, so all drug-outcome pairs are eligible to serve as negative controls. We chose to use the same drug-outcome pairs that were negative controls in real data for consistency in our experimental design. In addition to the simulated dataset without any injected signals (where all RR = 1), five databases were creating using signals injected at RR = 1.25, 1.50, 2, 4, and 10 for only the positive control test cases. For evaluation purposes, we restricted ourselves to drug-outcome pairs in the OSIM dataset with enough statistical power to detect a relative risk of 1.25 based on the age-by-gender-stratified drug and outcome prevalence estimates [14].

To gain insight into the ability of a method to distinguish between positive and negative controls, the effect estimates were used to compute the Area Under the receiver operator characteristics 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. We calculated AUCs for each parameter combination in each of the 6 simulated databases. We also evaluated the distribution of estimates for each method by comparing the point estimates to the true injected signal size. Mean squared error was computed as $MSE = mean\left((\log(RR) - \log(RR_{true}))^2\right)$.

3 Results

3.1 Simulation Validation

As Fig. 2 shows, both databases demonstrate the same proportion of males (40 %) and same age distribution (mean = 37.6, se = 17.7) across their respective populations. While both databases contain the same number of unique drugs and unique conditions per person, MSLR was observed to have a slightly higher number of total drug and condition records in the database. The scatter plot shows

	MSLR	OSIM2
% male	0.4	0.4
Age (mean)	37.6	37.6
Age (stdev)	17.7	17.7
Avg length of observation (days)	585.8	599.2
Avg # of drug records	13.3	11.9
Avg # of unique drugs / person	11.0	11.0
Avg # of condition records	17.7	17.5
Avg # of unique conditions / person	14.0	14.0

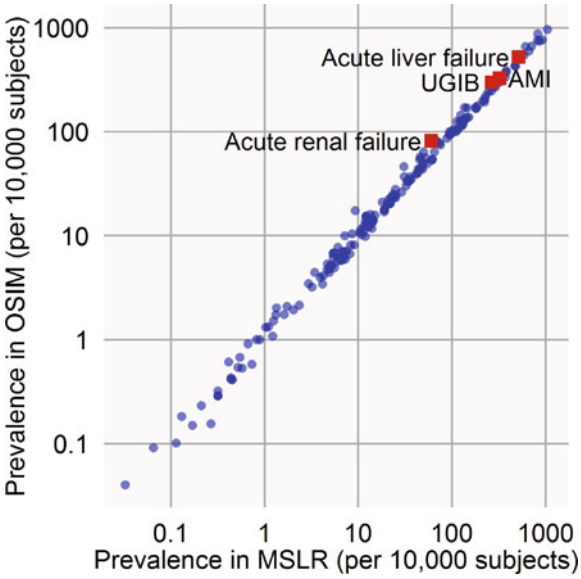

Fig. 2 Comparison of MSLR and OSIM2 population characteristics. The scatterplot shows the prevalence of all 178 drugs and 4 conditions encountered in the ground truth set. *AMI* acute myocardial infarction, *UGIB* upper gastrointestinal bleeding, *MSLR* MarketScan lab supplemental, *OSIM2* probabilistic framework

strong correlation in the prevalence of drugs and conditions between the two databases, which is consistent across both rare and commonly-occurring exposures.

Figure 3a illustrates modeling of confounding by age. Pioglitazone is a treatment for diabetes, not known to be associated with GI bleeding. However, unadjusted IRR (Incidence Rate Ratio) in MSLR yields a positive and statistically significant effect: 1.57 (1.38–1.79). One explanation for this effect is confounding by age, as patients with diabetes are older than the population in general, and the risk of UGIB increases with age. In MSLR, Mantel–Haenszel adjustment of age from stratified estimates within age deciles results in a non-significant effect: 1.08 (0.95–1.23). Repeating these analyses in OSIM2 demonstrates a similar pattern to MSLR: with an unadjusted estimate, we observe a positive effect IRR = 1.43 (1.37–1.51), which is almost fully attenuated with age adjustment, IRR = 1.09 (1.05–1.15). Figure 3b illustrates an example of confounding by indication by looking at the effect of sitagliptin and AMI. Sitagliptin is another antihyperglycemic treatment, but unlike some other drugs in the class, is not known to have a causal effect on AMI. However, its indication, diabetes, is a strong cardiovascular risk factor. When evaluating unadjusted IRR, both MSLR and OSIM show consistent, large positive effects. However, when imposing restriction by indication, we see the effect is attenuated as non-significant in OSIM, while MSLR suggests a negative association between sitagliptin exposure and AMI. This suggests OSIM may be modeling some of the confounding by indication, but may not fully model the channeling bias

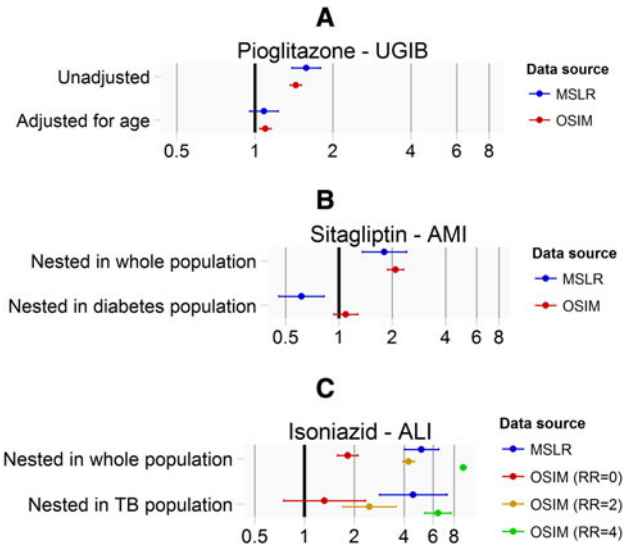

Fig. 3 Three examples of drug-outcome relationships in MSLR and OSIM; **a** pioglitazone and gastrointestinal bleeding, **b** sitagliptin and acute myocardial infarction, **c** isoniazid and acute liver injury. *MSLR* MarketScan lab supplemental, *OSIM2* probabilistic framework, *UGIB* upper gastrointestinal bleeding, *AMI* acute myocardial infarction, *ALI* acute liver injury

that exists in real-world practice. Figure 3c illustrates the modeling of the positive control association between isoniazid and acute liver injury. Isoniazid is an effective treatment for tuberculosis (TB), but is a known hepatotoxic agent. In MSLR, the unadjusted IRR = 5.09 (4.01–6.45), which is consistent when restricted to patients with

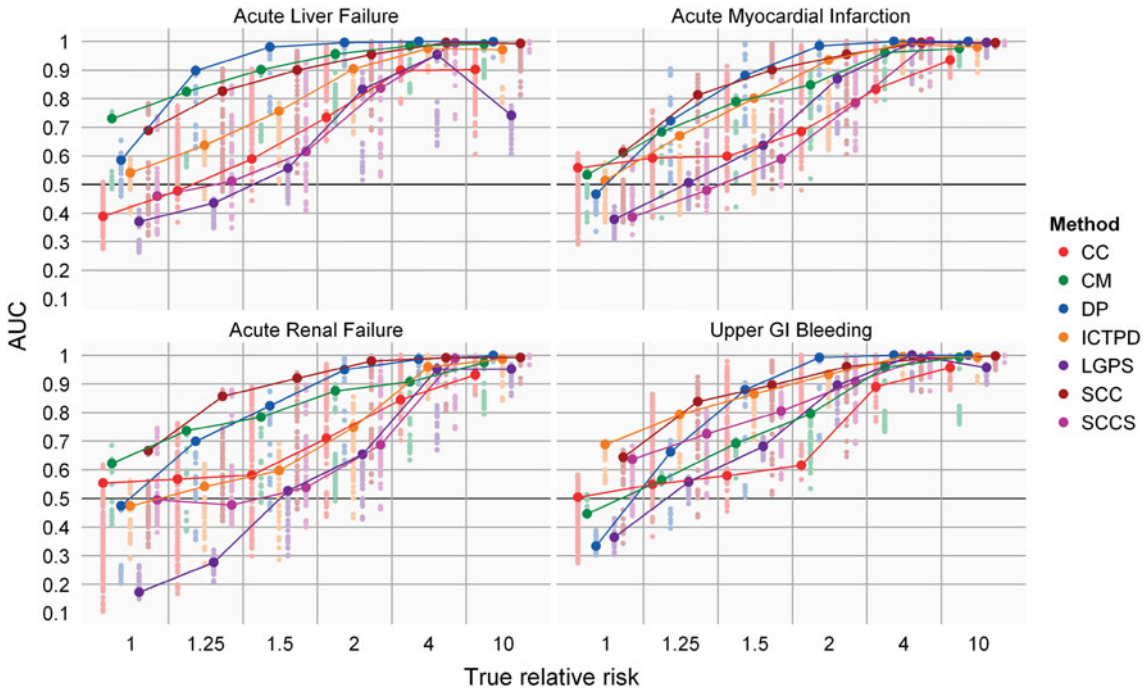

Fig. 4 Area under the receiver operator characteristics curve (AUC). *Light dots* indicate the performance of all different settings. *Dark colors* and *lines* indicate the settings that achieved the highest AUC in real data. Note that the *left*-to-*right* order of the method performance in this graph is the same as the *top*-to-*down* order in the legend. *CC* case-control, *CM* cohort method, *DP* disproportionality methods, *ICTPD* information component temporal pattern discovery, *LGPS* longitudinal gamma poisson shrinker, *SCC*self-controlled cohort, *SCCS* self-controlled case series

tuberculosis: IRR = 4.53 (2.83–7.24). In OSIM, prior to injecting signals, we see the unadjusted IRR illustrates residual positive bias, while the confidence interval for IRR restricted to TB patients contains RR = 1. We see that injecting signals in OSIM2 with RR = 2 and RR = 4 results in effect estimates that bound the observed estimate from MSLR.

3.2 Methods Performance

Figure 4 provides the predictive accuracy of each of the methods within the 4 outcomes, as function of the true relative risk that was injected as signals for the positive controls. The connected lines represent the single analysis within the method that was identified as having the highest predictive accuracy in real data [7–13]. The color dots in the background represent the performance of other analysis choices within each method. At RR = 1, we would expect predictive accuracy should be near AUC = 0.50, since there is nothing associated with the effect size that was injected to distinguish the positive controls from the negative controls. In general, we observe that the predictive accuracy of all methods increases as the effect size among the positive control increases. When the true relative risk for positive controls is greater than 4, all methods are highly predictive, with AUC > 0.90. In contrast, when $RR_{true} = 1.25$, no method is able to achieve AUC > 0.90. At $RR_{true} = 2.0$, there is generally at least one analysis of each method capable of near perfectly discriminating between positive and negative controls, but each method also has settings with AUC near 0.70. The optimal settings highlighted here vary in performance across the four outcomes, but three methods—new user cohort design, self-controlled cohort design, and disproportionality analysis—achieve AUC > 0.80 at $RR_{true} = 2.0$ in all scenarios.

Figure 5 illustrates the impact of signal injection at various sizes on the effect estimates for the four outcomes of interest. For each method, the settings producing the highest AUC in real data were used. Blue dots indicate negative controls, orange dots indicate positive controls at various levels of injected signal magnitude. Across the four outcomes, we observed all method to be increasingly negatively biased as the true effect size increased. SCCS was observed to have smaller variance in the estimate distribution, and demonstrated negative bias at all effect sizes larger than 1. Case–control design and LGPS was observed to be positively biased across all four outcomes when $RR_{true} < 2$. The new user cohort method (CM) is observed to have the largest variability in the estimate distribution across all methods.

Figure 6 shows the MSE as a function of the true relative risk that was injected as signals for the positive

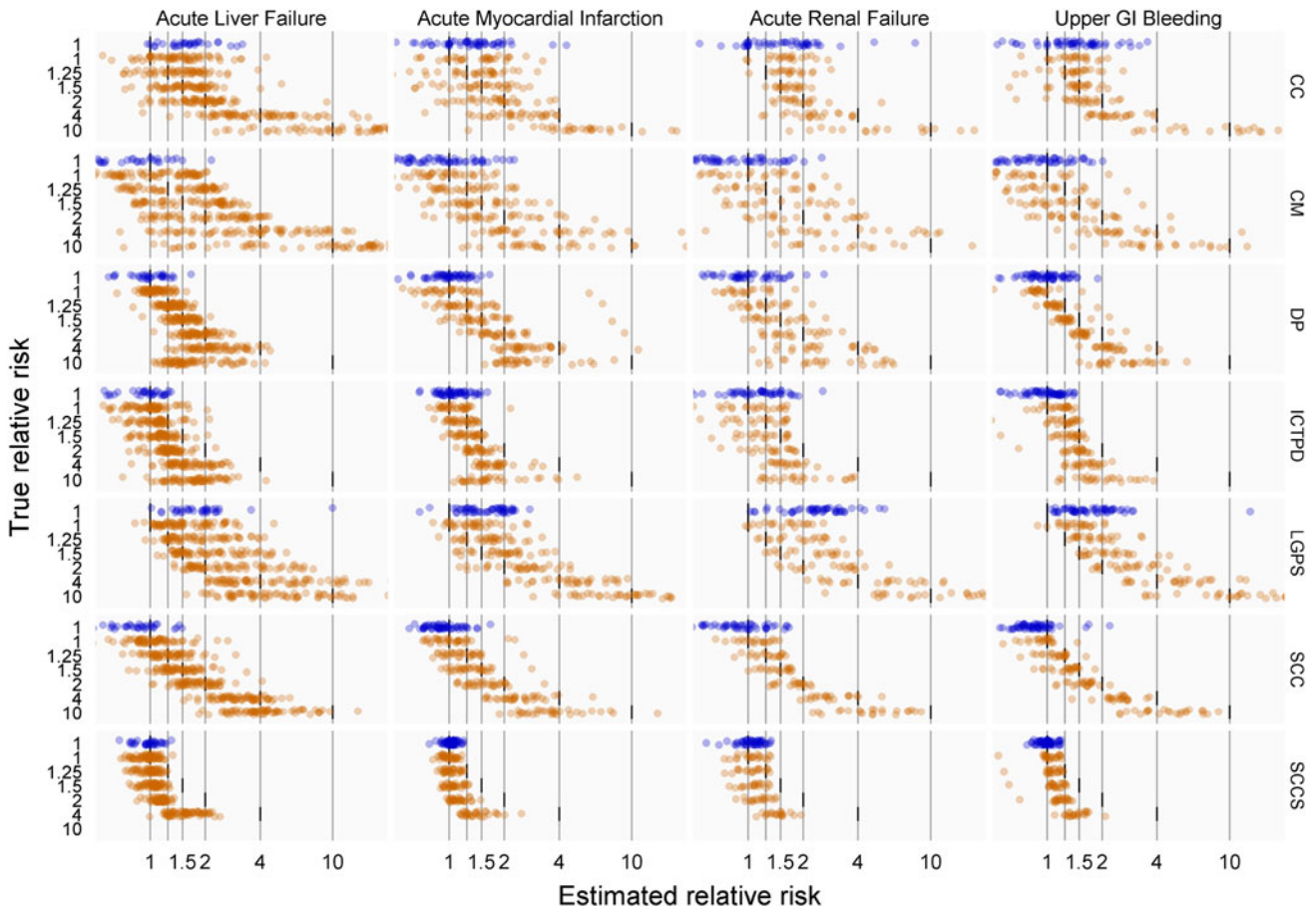

Fig. 5 Effect estimates for each method, by true relative risk. Each *dot* represents an effect estimate for a single drug-HOI combination. *Blue* indicates negative controls, *orange* indicates the positive controls with various magnitudes of injected signals. For each method, the settings producing the highest AUC in real data are used

controls. The connected lines represent the single analysis within the method that was identified as highest predictive accuracy in real data [7–13] and corresponds to the data shown in figure 5. The color dots in the background represent the performance of other parameter settings within each method. Most settings of the methods produce high MSE, and although the optimal settings for each method produce lower MSE for lower true RRs, every method has high MSE when the true RR is high.

4 Discussion

In this paper, we discuss a simulation framework that generates a simulated patient population and applies a Markov model to generate drug and condition records based on patterns found in real data. The objective is to produce a complete observational database in its relational form, with demographics, observation periods, and real-world values for drug dispensings and diagnosis codes linked to patient identifiers and longitudinally distributed. When structured in a common data model format, this simulated data is sufficiently comparable to real data in structure and content that analysis methods designed for use in the real data can be directly applied in simulation. This approach allows the complete analysis method, including all data manipulations required to restructure the data into an analytical dataset for subsequent statistical analysis, to be tested and evaluated across a wide array of drug-outcome scenarios. This modeling paradigm is quite different from other simulation approaches commonly used in methodological research, where an analytical dataset with a limited set of hypothetical variables (e.g. X, Y, Z, W) is generated repeatedly following a well-defined mathematical model to test a specific aspect of one particular type of bias [4, 5, 15, 16].

A primary motivation for developing this alternative simulation approach is the recognition that many potential sources of bias concurrently exist within an observational healthcare database, and the real-world performance of analysis methods needs to be evaluated in the context of all of those biases. With observational studies, evaluating the

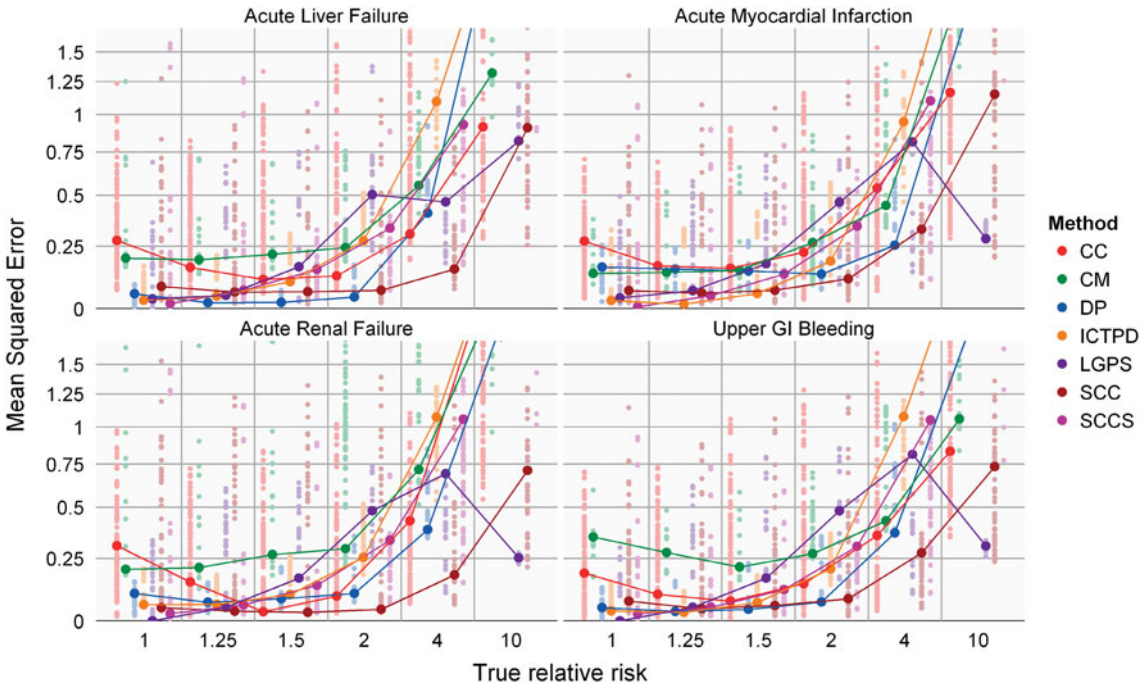

Fig. 6 Mean squared error as a function of injected signal size. *Light dots* indicate the performance of all different settings. *Dark colors* and *lines* indicate the settings that achieved the highest AUC in real data. Note that the *left*-to-*right* order of the method performance in this graph is the same as the *top*-to-*down* order in the legend. *CC* case-control, *CM* cohort method, *DP* disproportionality methods,*ICTPD* information component temporal pattern discovery *LGPS* longitudinal gamma poisson shrinker, *SCC* self-controlled cohort, *SCCS* self-controlled case series

performance of an analytical method always needs to be in the context of the data the method is applied to, as the interaction of the bias that exist in the data and the degree to which the method addresses those specific biases is ultimately what determines the accuracy of the effect estimate. One fundamental limitation of conducting methodological research in real data is the lack of a true gold standard that can be used to evaluate effect estimates produced by a method. In real data, the true effect size is never known with certainty, must be assumed or approximated, and may not be constant across different patient subpopulations.

Simulation offers a solution by enabling the researcher to define a true effect size and evaluate whether an analytical approach can recover the truth from the data. For methodological research purposes, we desire to have a simulation model with sufficient fidelity to properly model all of the complex interactions and temporal relationships between patient characteristics, health service behaviors, disease natural history, medical treatment utilization, and health outcomes. OSIM represents a step forward toward this ultimate goal. In the validation efforts presented here, we show that OSIM can accurately model a population that produces marginal summary statistics that are in line with the source data. We also demonstrate that aspects of bias observed in real data manifest in the simulation, such as confounding by age and indication. However, OSIM is a simplified model of reality and the limitations of the model should be considered when interpreting findings. Real data has a multitude of potential sources of bias that are not explicitly modeled in OSIM, including time-dependent confounding, informative dropout, and protopathic bias. The use of total condition count and total observation time as stratification factors in the model generation present a simple form of unmeasured confounding, but the model does not adequately represent the extent of unmeasured confounding from other factors commonly missing in observational data (like smoking, BMI, socioeconomic status) which chronically plague epidemiologic investigations. Only certain dependencies are modeled, and when injecting signals we do not consider differential misclassification of cases. Injected signals currently are constant across subpopulations, and have a constant hazard within a clearly delimited risk window. In reality, we know hazard functions likely vary over time and between patients, and can be variable depending on the drug-outcome scenario, with acute onset effects and latent events expressed under very different circumstances. Despite these limitations, we think simulation studies can complement real-world investigations and provide some evidence to support our understanding of the characteristics of analysis methods applied to observational data.

In this study, we evaluated the predictive accuracy of an array of alternative methods across different effect sizes. In general, we found that all methods demonstrate a similar pattern when measuring AUC. Namely, if the objective of a risk identification system is to identify large effects (RR $>$ 4), we can take some level of comfort that most methods can achieve this goal with high accuracy. There is considerable value from a public health perspective in having confidence that large safety effects will be identified reliably once a national surveillance system becomes operational. Conversely, if we hope that a risk identification system can discriminate subtle effects with RR $\leq$ 1.5 from observational data, we observe that all methods are less reliable in this endeavor with many methods yielding AUCs between 0.60 and 0.80. This suggests that while there is substantial evidence to be learned from observational analyses of small effects, results should be interpreted with caution and not be considered definitive findings. These findings in simulated data agree to some extent to what was observed in real data: self-controlled methods such as self-controlled cohort and ICTPD are better at discriminating positive from negative controls than the case–control method; methods exhibit large degree of variability in the error distribution; and the directionality of bias can vary within outcome and across methods. We also observed some discordant results when comparing real-world studies with in our simulation. For example, the analysis within the SCCS method which performed best in real data did not perform well as other analysis choices within the method. We posit this finding may be due to the optimal analysis in real data using multivariate adjustment for time-varying exposures to other drugs, but this type of time-dependent confounding is not modeled in OSIM so a univariate SCCS model may be preferred in this context. As another example, disproportionality analysis was observed to have better performance in simulation than in real data, but this may simply be an artifact of the simulation having fewer sources of confounding. These inconsistencies in findings raise a challenge for methodological research, in that it is difficult to determine which source of evidence is more reliable. Results from simulation are compelling in that the ground truth is well-defined, but the limitations of the simulation model represent the very challenges in real-world data that observational analyses are trying to overcome. On the contrary, empirical evaluations in real data are greatly limited by the integrity of the reference set developed; binary classification of positive controls and negative controls is often not sufficient but generally all that is possible for real-world investigations. Weighing evidence across real-world and simulation evaluations represents a tradeoff between one's comfort in the assumptions around the underlying data generation model versus assumptions around the unknown relationship between real drugs and outcomes, with the hope that the evidence is ultimately complementary toward a common goal.

The magnitude of bias that persists in the analysis is such that methods cannot always properly distinguish between null effects from positive effects in this setting. The range of estimates across each of the methods illustrates both the mean bias and variability of error that can be observed. While each method exhibits a different error profile, it is notable that all methods demonstrate sufficient variability around the estimates that negative controls with RR $=$ 1 cannot be readily distinguished from positive controls with RR $=$ 2. While much discussion often centers on reducing bias is a primary motivation in the design of observational studies, these findings suggest that developing strategies for reducing the variability in the error distribution may also provide important gains in the improving the quality of evidence that can be derived from these data sources. To an extent, these findings reinforce the intuition that effect estimates from observational studies with RR $<$ 2 are generally considered less reliable and require greater scrutiny [17]. A challenge we face is that many of the notable safety issues in the past decade, such as the cardiovascular risks of rofecoxib and rosiglitazone, have been associations with effects closer to RR $=$ 1.5 than RR $=$ 4. One opportunity for the use of a simulation model like OSIM2 is to facilitate the research community in developing better analytical methods that have improved performance characteristics in setting with small effects, such that observational evidence can be interpreted with greater confidence.

5 Conclusion

Simulation offers the potential to facilitate the methodological research necessary to develop a risk identification system. OSIM2 offers an approach for modeling observational data and reflects many characteristics of disease and treatment utilization, but does not fully represent the array of confounding that persists in real-world studies. Simulation results confirm findings from studies on real data which suggest observational analyses can achieve strong predictive accuracy but suffer from variable bias and large error in effect estimation. Evidence from simulation supports the belief that observational data can be used to identify large risks with RR $>$ 2, but highlights the challenges that should be expected when trying to detect smaller effects.

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. This work was supported by UBC/ProSanos for implementing OSIM2 through funding by FNIH. The authors thank Susan Gruber for the OSIM2 assessment she shared with the OMOP Statistics working group.

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. Ryan PB, Madigan D, Stang PE, Marc Overhage J, Racoosin JA, Hartzema AG, et al. 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.
2. 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;50(10):890–7.
3. Murray RE, Ryan PB, Reisinger SJ. Design and validation of a data simulation model for longitudinal healthcare data. AMIA Annu Symp Proc. 2011;2011:1176–85.
4. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. Am J Epidemiol. 2007;166(6):646–55.
5. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf. 2008;17(6):546–55.
6. 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.
7. Madigan D, Schuemie MJ, Ryan PB. Empirical performance of the case–control design: lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0105-z.
8. 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.
9. 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.
10. Norén GN, Bergvall T, Ryan PB, Juhlin K, Schuemie MJ, Madigan 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. 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.
12. Ryan PB, Schuemie MJ, Madigan D. Empirical performance of the self-controlled cohort design: lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0101-3.
13. 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.
14. Armstrong B. A simple estimator of minimum detectable relative risk, sample size, or power in cohort studies. Am J Epidemiol. 1987;126(2):356–8.
15. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol. 2011;174(11):1213–22.
16. Liu W, Brookhart MA, Schneeweiss S, Mi X, Setoguchi S. Implications of M bias in epidemiologic studies: a simulation study. Am J Epidemiol. 2012;176(10):938–48.
17. Temple R. Meta-analysis and epidemiologic studies in drug development and postmarketing surveillance. JAMA. 1999;281(9):841–4.