Drug Saf (2013) 36 (Suppl 1):S59–S72
DOI 10.1007/s40264-013-0099-6

ORIGINAL RESEARCH ARTICLE

Empirical Performance of a New User Cohort Method: Lessons for Developing a Risk Identification and Analysis System

Patrick B. Ryan · Martijn J. Schuemie · Susan Gruber · Ivan Zorych · David Madigan

Abstract

Background Observational healthcare data offer the potential to enable identification of risks of medical products, but appropriate methodology has not yet been defined. The new user cohort method, which compares the post-exposure rate among the target drug to a referent comparator group, is the prevailing approach for many pharmacoepidemiology evaluations and has been proposed as a promising approach for risk identification but its performance in this context has not been fully assessed.

Objectives To evaluate the performance of the new user cohort method as a tool for risk identification in observational healthcare data.

Research Design The method was applied 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) and in 6 simulated datasets with no effect and injected relative risks of 1.25, 1.5, 2, 4, and 10, respectively.

Measures Method performance was evaluated through Area Under ROC Curve (AUC), bias, and coverage probability.

Results The new user cohort method achieved modest predictive accuracy across the outcomes and databases under study, with the top-performing analysis near AUC >0.70 in most scenarios. The performance of the method was particularly sensitive to the choice of comparator population. For almost all drug-outcome pairs there was a large difference, either positive or negative, between the true effect size and the estimate produced by the method,

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,
Titusville, NJ, USA

P. B. Ryan (✉)
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

S. Gruber
Harvard School of Public Health, Cambridge, MA, USA

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

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

although this error was near zero on average. Simulation studies showed that in the majority of cases, the true effect estimate was not within the 95 % confidence interval produced by the method.

Conclusion The new user cohort method can contribute useful information toward a risk identification system, but should not be considered definitive evidence given the degree of error observed within the effect estimates. Careful consideration of the comparator selection and appropriate calibration of the effect estimates is required in order to properly interpret study findings.

1 Background

Pharmacovigilance encompasses all scientific and data gathering activities relating to the detection, assessment, and understanding of adverse events of medical products through the product lifecycle. These activities principally involve the identification and evaluation of safety signals, which refers to "a concern about an excess of adverse events compared with what is expected to be associated with a product's use [1]." This assessment of the difference between observed and expected can be conceived in the ideal as a relative comparison between the effect among those patients exposed to the treatment of interest with the counterfactual of the same persons having not been exposed to the treatment. The challenge with this ideal is that the counterfactual cannot be directly observed, but instead represents a hypothetical alternative [2]. Randomized, placebo-controlled trials can provide a robust estimate of these relative risks, as randomization minimizes the risk that the treated and placebo cohorts are systematically different in some way that would bias the comparison, and the placebo serves as a proxy for the counterfactual of the exposed population having been alternatively 'unexposed' [3]. In this regard, association measures from randomized control trials (RCTs) can be interpreted as true causal effect measures, since randomization ensures the study preserves the exchangeability assumption needed to evaluate the counterfactual [4, 5]. RCTs serve as the primary information source prior to approval, and can often offer the most definitive evidence for evaluating adverse events that occur frequently enough within the initial time exposed. However, even large RCTs are underpowered for rare events and may not capture data for the full time-at-risk needed to assess long-term effects [6]. Once a product is approved, safety data collection and risk assessment based on observational data become critical for evaluating and characterizing a product's risk profile and for making informed decisions on risk minimization [1].

A prevailing approach to observational studies of drug safety issues is the cohort design, where patient exposed to the target drug of interest are compared with patients belonging to some referent group, typically either patients not exposed to the target drug or, more narrowly, those exposed to a specific alternative treatment or 'active comparator'. The design is conceptually similar to that of the randomized trial in that two groups are followed prospectively from initiation of treatment and differences in rates of outcomes can be measured. In this regard, the referent group in the observational cohort design is serving as the analogue to the placebo arm of RCT, and is intended to represent patients who counterfactually may have otherwise received the target treatment but instead receive an alternative treatment that does not have an effect on the outcome. In contrast to RCTs, where treatment assignment is driven by the randomization scheme of the investigator, observational cohort studies require an assumption that treatment assignment is ignorable, after controlling for recorded confounders [7]. The conceptual arguments for the use of the new user cohort design in pharmacoepidemiology have been widely discussed [8–11]. By identifying patients who start a course of treatment with the target medication, and use the initiation of therapy as the start of follow-up, the new user cohort design models the behavior of a RCT where treatment commences at the index study visit. Defining the index date at start of treatment allows a clear separation of baseline characteristics, which occur prior to index date and can be used as covariates in analysis without concerns of inadvertently introducing intermediate variables that arise between exposure and outcome. Excluding prevalent users as those without sufficient washout period prior to first occurrence of exposure allows a truer estimation of the time-to-event relationship between drug start and outcome incidence by reducing bias due to depletion of susceptibles.

The new user cohort approach has been advocated as the primary design to be considered for studies of drug safety [12, 13] and comparative effectiveness [14]. Several aspects of design decisions within the new user cohort approach, such as time-at-risk definition [15, 16], covariate adjustment through propensity scoring [17–21], and propensity score trimming [22], have been illustrated through selected case studies. As with most observational study designs, there is little empirical evidence that quantifies the overall performance of the new user cohort design and the impact of decisions within the cohort design when applied to real-world observational healthcare databases across an array of drug-outcome scenarios. In this study, we tested an implementation of the new user cohort method in 5 real observational healthcare databases and 6 simulated datasets, retrospectively studying the predictive accuracy of the method when applied to a collection of 399 drug outcome pairs-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 method can be expected to identify true effects and discriminate from false findings and explore the statistical properties of the estimates that the new user cohort method generates.

2 Methods

We conducted the methodological research experiment against five disparate observational healthcare databases to allow evaluation of performance across different populations and data capture processes: 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 Healthcare Centricity™ (GE, 11.2 m persons) database. GE is an electronic health record (EHR) database, the other four databases contain administrative claims data. A 10 m-person simulated dataset was also constructed using the OSIM2 simulator [23] to model the MSLR database, and replicated 6 times to allow for injection of signals of known size (relative risk = 1 (no effect), 1.25, 1.5, 2, 4, 10). The data used is described in more detail elsewhere [24].

All databases were standardized into the OMOP common data model format [25]. This allowed a single implementation of the new user cohort method, parameterized to allow systematic manipulation of individual analysis choices (such as the definition of time-at-risk or the covariate selection algorithm) and consistent application of the method across the data network. The source code and full specifications for the new user cohort method are available at: http://omop.org/MethodsLibrary. Briefly, the program creates two cohorts, and evaluates the rate of events observed within a defined time window following the cohort index date using logistic regression for the outcome model. The first cohort is comprised of patients who had a record indicating at least one exposure to the target drug, where the index date is defined as the date of first exposure. Exposure is inferred from available data elements in the source database, including pharmacy dispensings, procedural administrations, and medication history records. The second cohort is comprised of those patients satisfying the criteria for the comparator group.

We tested five alternative comparator group definitions. First, we define one active comparator drug for each target drug, requiring that the comparator be the most prevalent treatment that shares the same indication as the target drug but falls within a different therapeutic class. For example, lisinopril is an angiotensin-converting-enzyme [26] inhibitor indicated for hypertension, and its selected comparator is amlodipine, which is a calcium channel blocker that is the most prevalent antihypertensive treatment outside of ACE inhibitor class. Second, we defined a group of active comparator drugs for each target drug, based on all treatments which share the same indication as the target drug but falling within a different therapeutic class. Continuing the example, the comparator drugs for lisinopril would be all antihypertensive drugs not in the ACE inhibitor class, including Angiotensin Receptor Blockers, calcium channel blockers, beta blockers and diuretics. Third, we defined a set of active comparator drugs which are known to be unrelated to the outcome of interest, including all ‘negative controls’ drugs used with test cases (‘negative control’ comparator group), and further restricted the comparator cohort to patients with a diagnosis of the indication of the target drug. In these three cases, the comparator cohort index date is defined as the date of first exposure to the active comparator. Fourth, we defined the comparator cohort as patients not exposed to the target drug but having a diagnosis record for the target drug indication, and use the date of first diagnosis as the index date. In the example, patients exposed to lisinopril would be compared with all patients diagnosed with hypertension who did not have any exposure to lisinopril. Finally, we use the entire database as a comparator group.

Both cohorts are restricted to patients who have a defined ‘washout period’ of observed unexposed time prior to the cohort index date, in order to increase confidence that the first exposure record truly reflects an incident user. In this study, the ‘washout period’ was defined as 180 days prior to first exposure. It is worth noting that if ‘washout period’ is set to 0 days, the method would include prevalent users and should no longer be considered a new user design. We executed one analysis using the 0-day washout window to facilitate comparison of the incident vs. prevalent user designs.

Both cohorts can also be optionally restricted to patients who have at least one diagnosis record for the target drug indication, and were tested with and without restriction. Patients who satisfy the criteria for both the target and comparator cohorts and have overlapping time-at-risk are excluded from both cohorts. With the cohorts established, covariate adjustment can be applied to minimize imbalance between the populations at baseline. The new user cohort design we have implemented estimates the propensity score [27] through logistic regression, then optionally restricts the cohorts through propensity score trimming and adjusts the effect estimates either through propensity score stratification or by using the propensity score as covariates in the outcome model. The propensity score model was tested using three alternative covariate selection approaches: the high-dimensional propensity score algorithm proposed by Schneeweiss et al. [17], which selects a defined number of top covariates related to exposure and outcome; an implementation by Brookhart et al. [28], that selects a set of covariates based on prevalence and association with exposure only; and a large-scale Bayesian regression approach [29] that uses all available covariates representing baseline conditions, drug exposures, and procedures to classify exposure status. For each of these covariate selection procedures, we use a covariate risk window prior to the cohort start date to ensure no intermediate variables are included in the propensity score model. Because these procedures are executed independently for each drug-outcome pair on each database, the specific covariates selected vary within each drug-outcome-database analysis. The specific values tested for each of the 10 analysis choices within the new user cohort method implementation are illustrated in Fig. 1.

In total, 126 different combinations of analysis choices were executed within this experiment against all observational databases [30], and each analysis was applied to 399 different drug-outcome pairs to generate an effect estimate and standard error for each pair. 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 [31]. For every database we restricted the evaluation 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 [32]. The entire dataset that contains all effect estimates and standard errors from all new user cohort experiments are available for download at: http://omop.org/Research.

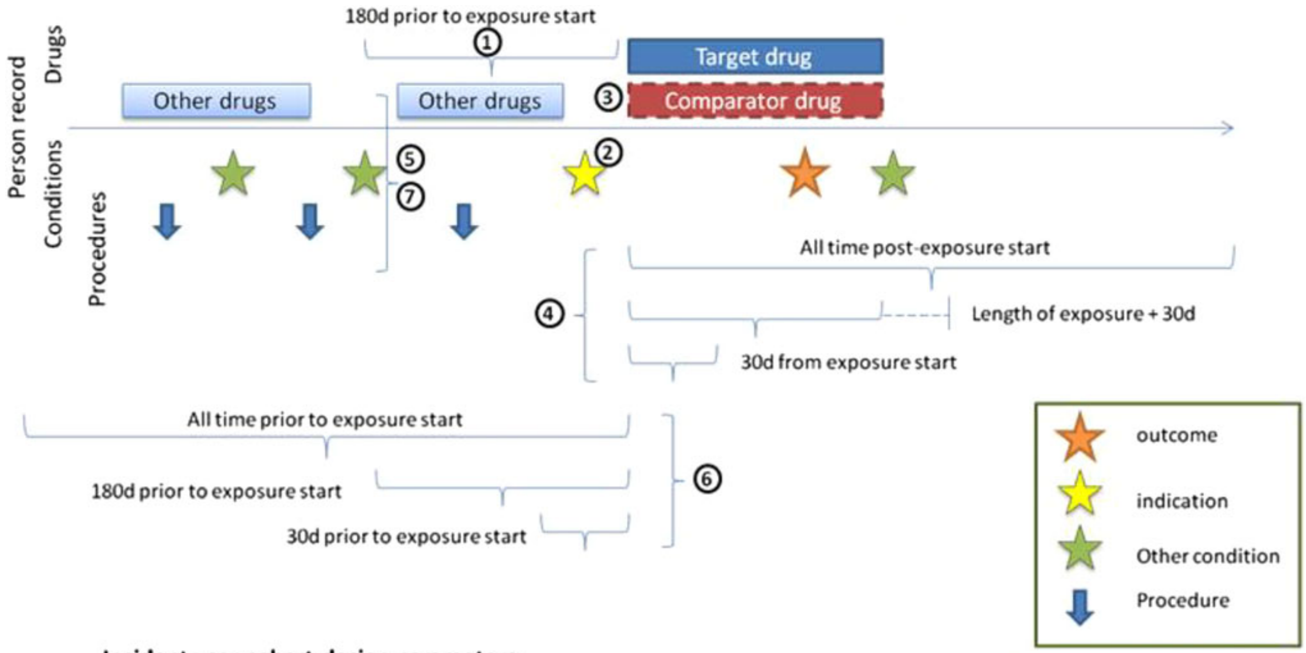
The assessment of the method validity requires a gold standard to be used to compare the methods estimates with some notion of a ground truth [9]. In this experiment, we assess multiple measures of methods performance by applying three strategies for ground truth. First, we use the dichotomous classification of our test cases as ‘positive controls’ and ‘negative controls’ to serve as a binary marker that can be used in assessing predictive accuracy. Second, for our ‘negative controls’, where we have no evidence to suggest that the drug is associated positively or negatively with the outcome, we can assume that the true relative risk = 1. Note, we can make no assumptions of the true relative risk for positive controls, as we only have evidence to suggest RR >1 but are unsure of the exact magnitude of effect. Instead, as a third approach, we inject signals of known sizes into simulated data for our ‘positive controls’, providing a defined measure of relative risk for which to compare the methods estimates.

With the binary classification of ground truth, we rank-order effect estimates across the test cases to compute the Area Under the receiver operator characteristics Curve (AUC) [33]. An AUC of 1 indicates a perfect prediction, which in this context means complete discrimination between the positive and negative controls at some threshold for the effect estimate. An AUC of 0.5 is equivalent to random guessing. As a frame of reference, the predictive accuracy of diagnostic tests commonly used in clinical practice, such as prostate-specific antigen screening for prostate cancer, mammography for breast cancer, and rapid strep tests have been evaluated using AUC, and range from 0.74–0.80 [34–36].

With ground truth established as a real-valued effect size ($\mathrm{RR}_{\mathrm{true}}$) for each test case, we can compare effect estimates ($\mathrm{RR}_{\mathrm{est}}$) with the true value to measure error. The error distribution across test cases can then be used to evaluate bias, which is simply the expected value of the error distribution ($bias = mean(\log(RR_{est}) - \log(RR_{true}))$). We also calculate mean squared error (MSE), as $MSE = mean\left((\log(RR_{est}) - \log(RR_{true}))^2\right)$. Bias and MSE are metrics that characterize the behavior of a method’s point estimate, but do not consider the behavior of the method’s measure of standard error. We also calculate coverage probability as the probability that an estimated confidence interval contains $\mathrm{RR}_{\mathrm{true}}$. Assuming nominal characteristics and an unbiased estimator, we expect a 95 % confidence interval to have a 95 % coverage probability. We assessed the consistency of coverage probability estimates among positive controls at 6 different $\mathrm{RR}_{\mathrm{true}}$ values from the simulated datasets.

3 Results

Figure 2 highlights the predictive accuracy, as measured by AUC, of all analyses across the four outcomes and four databases. The majority of analyses were executed with the comparator group defined as a single drug with the same indication. In some circumstances, specific settings of the method achieved moderate predictive accuracy, with AUC >0.65 in at least one database for acute liver injury, acute renal failure, and gastrointestinal bleeding. However, many analyses with this comparator produced low predictive accuracy, with AUC near 0.50. Expanding the active comparator group to include all drugs with the same indication as the target drug did not substantially change predictive accuracy. The highest levels of predictive accuracy were achieved when using an active comparator group defined by ‘negative control’ drugs not associated with the outcome, which were consumed among patients with the



Incident user cohort design parameters:

1. **Required observation time prior to exposure:** 180d, None
2. **Nesting within population with the indication of the target drug:** Yes, No
3. **Comparator population:** 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, 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
4. **Time-at-risk:** *Length of exposure + 30d*, 30d from exposure start, All time post-exposure start
5. **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
6. **Covariate eligibility window:** 30d prior to exposure, 180d prior to exposure, All time prior to exposure
7. **Dimensions to include as potential covariates:** Drugs only, drugs and conditions, drugs and conditions and procedures
8. **Additional covariates in propensity score model:** Age, sex, index year, Charlson index, number of drugs, number of visits, number of procedures
9. **Propensity score trimming:** None, Trim lower 5% from the comparator group and the upper 5% from the target group
10. **Metric:** Propensity score stratification using Mantel Haenszel adjustment over 5 strata, Propensity score stratification using Mantel Haenszel adjustment over 20 strata, 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 adjustment using propensity score as continuous variable in logistic regression outcome model, Unadjusted odds ratio from univariate logistic regression predicting outcome from exposure

Fig. 1 Parameters within incident user cohort design

indication for the target drug. This highlights one drawback of using an active comparator group. When the risk of the outcome is increased more by the use of the comparator drug than by the test drug under consideration RR_{TRUE} is <1, masking the fact that the test drug also increases risk for the outcome. The use of the 'negative control' comparator was observed to have AUC >0.70 in most database-outcome scenarios. In MSLR, for acute liver injury and gastrointestinal bleeding, using a single active comparator drug yielded a higher AUC than the 'negative control' comparator group which was optimal in all other situations.

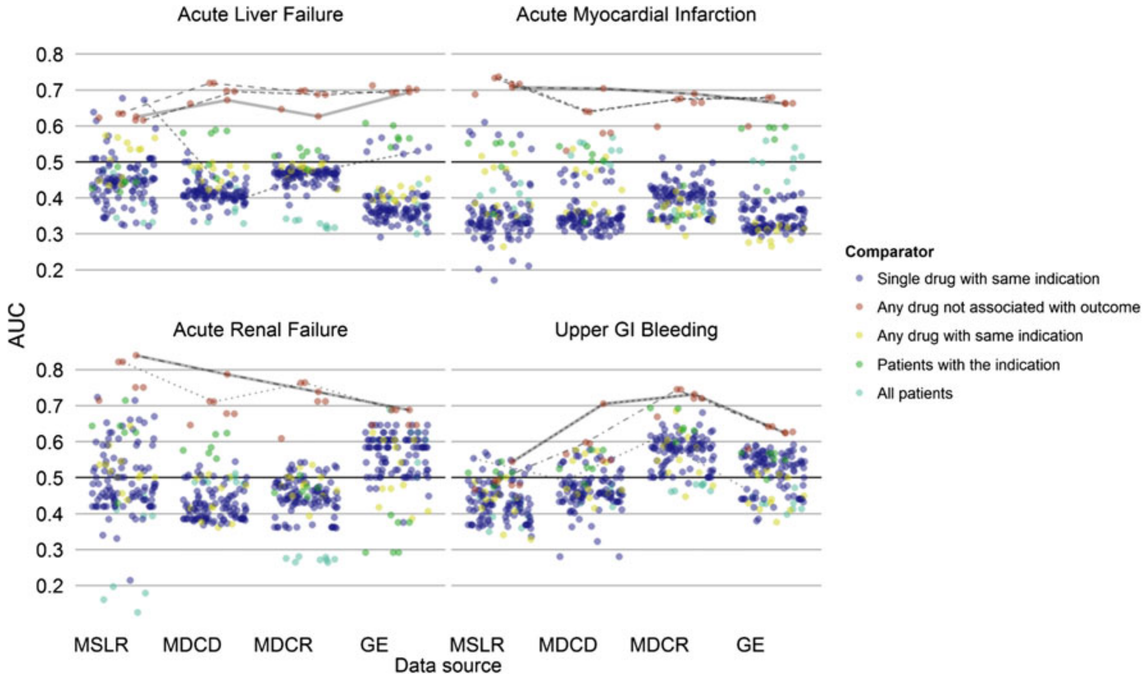


Fig. 2 Area under ROC Curve (AUC) for cohort parameters, by outcome and database. *Each dot* represents one of the 126 unique parameter combinations of the incident user cohort design. The *solid grey line* highlights the parameter that had the highest average AUC across all 20 outcome-database scenarios. The *dashed lines* identify each setting with the highest AUC for each database within each outcome. The *dots* are colored by the choice of comparator used in the analysis. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

The analysis that yielded the highest average AUC across the drug-outcome scenarios was CM:21000216, which used the 'negative control' comparator group, applied a 180d washout period to define incident users, and defined the time-at-risk as the period from the start of exposure through 30 days after the end of exposure. This analysis fit a model with Bayesian logistic regression using all available covariates defined by looking in the 180d prior to the cohort entry for conditions, drugs, and procedures, plus covariates for age, sex, index year, Charlson index [37], number of drugs, number of visits, and number of procedures to estimate the propensity score, and then used the propensity score as a covariate in the logistic regression for the final outcome model.

The choice of comparator group was the primary discriminator of predictive accuracy across databases and outcomes. Among the notable parameters that were directly compared head-to-head, the choice of covariate selection algorithm and the manner in which the propensity score was applied in the final model (e.g. stratification or covariate adjustment) for did not materially impact the predictive accuracy estimates.

"Appendix" contains the effect estimates for all test cases across the five databases using this optimal analysis (CM:21000216). The impact of choice of comparator is illustrated using the four specific test cases for acute liver injury as shown in Fig. 3.

Both lisinopril and ciprofloxacin are considered positive controls for acute liver injury based on product labeling and systematic review [38]. Ciprofloxacin is a fluoroquinolone antibiotic, and whose single active comparator was identified as amoxicillin. When evaluating the risk of acute liver injury on ciprofloxacin, we observe that three comparisons ('all patients', 'any drug with same indication', and 'single drug with same indication') produce positive, statistically significant associations consistent with our expectations for this positive control. However, when comparing ciprofloxacin with the 'negative control' cohort or against other patients with infections that are not prescribed ciprofloxacin, we find a negative association. A possible explanation for this inconsistency is that patients with infections that are not commonly treated by a fluoroquinolone, such as community-acquired pneumonia, may be associated with patients who are sicker and therefore have a higher baseline risk for liver injury than those on ciprofloxacin. In the case of lisinopril, only the comparison with the 'negative control' comparator group yielded a positive association (RR = 2). Other comparison choices, including the comparison to the single drug amlodipine, which is not known to cause acute liver injury, showed no effect.

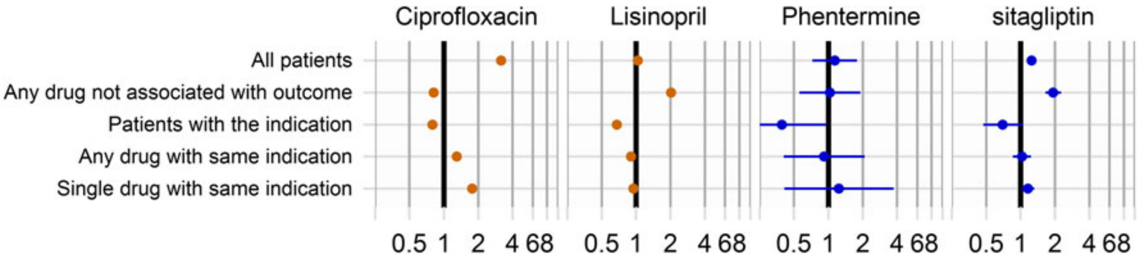


Fig. 3 Relative risk and 95 % confidence interval for 4 example drugs and acute liver injury, across comparators. *RR* relative risk, *Blue* negative controls; *Orange* positive controls; *each line* represents point estimate and 95 % confidence interval for the drug-outcome pair in a particular comparator

Two negative controls are also presented: sitagliptin is an oral antihyperglycemic treatment which can be compared with pioglitazone, or more broadly with any antidiabetic treatment that is not a DPP-4 inhibitor. In the sitagliptin example, the 'negative control' comparator group produced a positive effect estimate when no such association was expected. This finding may be explained in part by detection bias due to differences is health service utilization immediately preceding sitagliptin initiation, which typically would involve a physician visit and regular chemistry profiles to monitor HbA1c levels, which would be less commonly observed among patients initiating therapy for any of the negative control drugs. The pattern of estimates observed across comparators with the positive control lisinopril is consistent with the pattern observed for the negative control sitagliptin, demonstrating the challenge in using the estimates to discriminate between positive and negative findings. Phentermine is used to treat obesity, and was compared with sibutramine. The negative control phentermine highlights an example where all estimates consistent find no association where none is anticipated.

Figure 4 shows the range of estimates observed for negative control test cases within the MDCR database for each of the comparator groups. If we assume negative controls have $\mathrm{RR}_{\mathrm{true}} = 1$, we would expect the distribution of effect estimates to be tightly centered around 1. When using 'all patients' as the comparator group, we observe strong positive bias, with almost all $\mathrm{RR}_{\mathrm{est}} > 1$. This may be the result of a high proportion of healthy patients not exposed to any drugs and periods of observation without health service utilization in the database, which may be systematically less likely to observe events than during the period immediately following exposure to any target drug. In contrast, when using 'patients with the indication' as a comparator group, we observed a negative bias, with most $\mathrm{RR}_{\mathrm{est}} < 1$. This may also reflect a bias in establishing an index date not based on treatment initiation, as patients potentially have a higher risk of having subsequent outcomes recorded following initial diagnosis of a disease (a result of a physician visit) than they do following a pharmacy dispensing (that may not have involved a physician visit).

For all three comparator groups that are defined by new user of an active comparator, the estimate distributions are centered more around one for all four outcomes. The error distributions for a comparator defined by a single drug with the same indication or all drugs with the same indication have a consistent mean and variance. The 'negative control' comparator group also demonstrates a small average error, suggesting no strong positive or negative bias. However, in all instances of active comparator groups, the variance in the error distribution is large, with observed estimates ranging from smaller than $\mathrm{RR}_{\mathrm{est}} = 0.50$ to larger than $\mathrm{RR}_{\mathrm{est}} = 2.00$. Since many test cases have large sample sizes which result in small standard errors, the consequence of the variability in the estimates is that many negative controls yield statistically significant effects (either positive or negative), even when no such effects should be observed [39]. Figure 5 shows the same analysis in the simulated dataset. As can be seen, the distribution of estimates is very similar to those in the real data. Some bias may be expected because the simulated data contains unmeasured confounding by baseline covariates, just as we would expect in real data.

Figure 6 shows the coverage probabilities on simulated data for the overall optimal design, CM:21000216. When no signals were injected and $\mathrm{RR}_{\mathrm{true}} = 1$, the coverage probability ranged from 26 % for gastrointestinal bleeding to 55 % for acute myocardial infarction. The coverage probabilities decreased in each outcome as the $\mathrm{RR}_{\mathrm{true}}$ increased, and was observed to be <10 % when $\mathrm{RR}_{\mathrm{true}} = 10$. For gastrointestinal bleeding, the true effects were larger than the upper bound of the estimated confidence intervals in the majority of the test cases. For the other three outcomes, it appeared equally likely that the estimated confidence interval could underestimate or overestimate the true effect size. This finding is consistent with the observation from the error distribution, which suggests that while the mean error is small, the variance in the error distribution is sufficiently large than any given estimate can be notably far from its expected value.

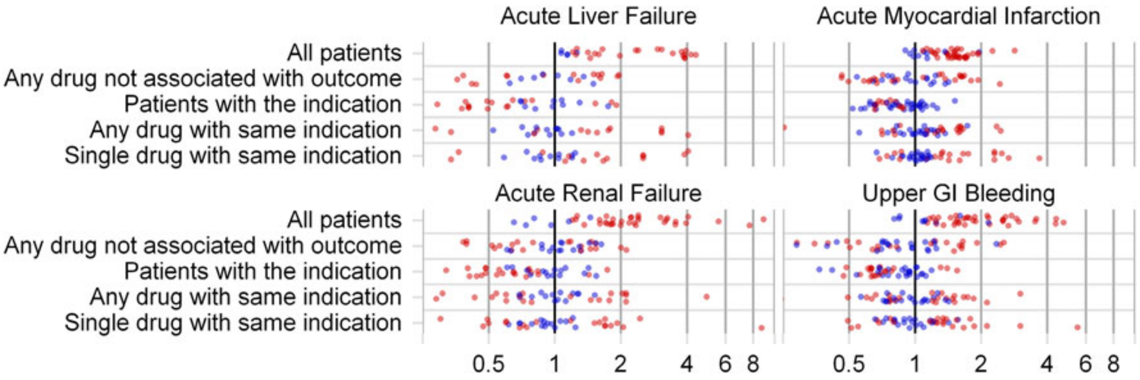


Fig. 4 Distribution of risk estimates across outcomes and comparators in the MDCR database. *Blue* indicates estimates that are not statistically significant ($p > 0.05$), *red* indicates estimates statistically significantly different from 1

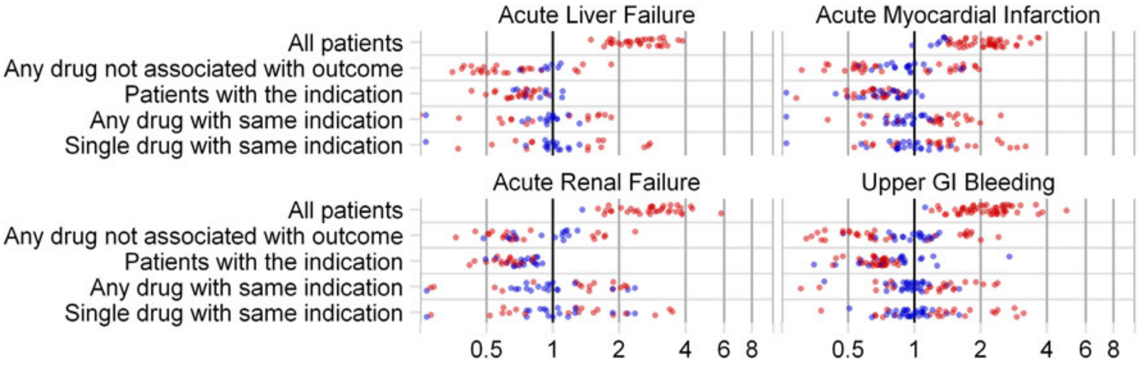


Fig. 5 Distribution of risk estimates across outcomes and comparators in the simulated database. *Blue* indicates estimates that are not statistically significant ($p > 0.05$), *red* indicates estimates statistically significantly different from 1

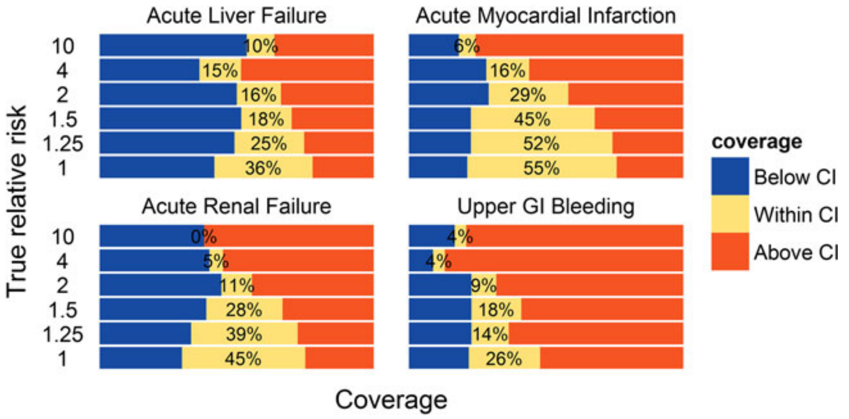


Fig. 6 Coverage probability of incident user cohort design in simulation at different levels of true effect size, by outcome. *CI* confidence interval

4 Discussion

In this paper, we evaluate the absolute performance of the new user cohort method in terms of predictive accuracy, bias, and coverage probability. The relative performance of the cohort method compared with other types of designs, such as case-control and self-controlled case series, is described elsewhere [24]. The cohort design is one of the most often-used in drug safety research using observational data, and to date, it has been the only methodological approach executed within the pilot projects of the Sentinel Initiative. Mini-Sentinel has applied multiple different approaches within the cohort design to assess potential safety risks, each using an active comparator drug to compare with the exposed population of interest. It conducted a fully unadjusted cohort study that evaluated the risk of gastrointestinal and intracranial hemorrhage among patients exposed to dabigatran by comparing them with patients exposed to warfarin [40]. It evaluated the risk of acute myocardial infarction (AMI) for varenicline by comparing the exposed population with patients exposed to bupropion, and performing minimal adjustment through stratification by age, gender, and plan [41]. It has proposed to conduct a customized propensity score-adjusted evaluation of saxagliptin and the risk of AMI through comparisons with sitagliptin, pioglitazone, long-acting insulins, and second-generation sulfonylureas [42]. Members of the Mini-Sentinel project have advocated for the use high-dimensional propensity score as an automated algorithm for covariate adjustment that can be applied within a new user cohort design, with the target drug contrasted with an active comparator with the same indication [43]. Additionally, the Sentinel Federal Partners Collaboration conducted two cohort analyses, one that assessed neuropsychiatric effects of oseltamivir, zanamivir, and amantadine, relative to ampicillin, azithromycin, and trimethoprim-sulfamethoxazole, and another which assessed the risk of heart failure of dronedarone by comparing with amiodarone [44]. The reason for the popularity of the new user cohort design and its leading role in Mini-Sentinel appear to be the notion that it "is a broadly applicable design that is fairly robust against investigator error" [12], and has been likened to RCTs. In a previous study, a cohort design analysis using propensity scores for a single drug-outcome pair was found to produce comparable estimates to several RCTs [9].

In our evaluation, we sought to systematically evaluate this method on a broader scale, but were unable to construct a gold standard of effect size estimates that were consistently confirmed by multiple RCTs. Therefore we chose to create a gold standard with just a binary classification: drug-outcome pairs that represent known adverse drug reactions and pairs where no causal relationship is believed to exist. Instead of evaluating whether the cohort design produces correct absolute effect size estimates, we evaluated this method on a simpler task, only requiring the estimates of the positive controls to be higher than those of the negative controls. We used the AUC as a metric of the performance on this task, and found many analysis choices within the cohort framework produced poor performance with AUCs around 0.5, meaning these cohort analyses could not reliably distinguish between positive and negative controls. Only when using a comparator group defined by 'negative control' drugs did we see an increase in performance, which is a strategy that has not often been used in scientific literature.

Even though we do not know the true effect sizes for the positive controls, we can assume RR = 1 for negative controls. For this subset we observed that, although the distributions of effect estimates are centered on 1 when using an active comparator, there is large variability with estimates ranging from 0.5 to 4.0. One possible explanation for these large error distributions could have been the fact that some of the selected comparators have a causal relationship with the outcome. However, we were able to test this potential hypothesis in simulated data where the comparator drugs were known not to have an effect on the outcome, and we found a similar error distribution with large variance suggesting that this explanation cannot fully explain the observed phenomenon. Instead, we hypothesize that even though the use of active comparators with the same indication and matching on propensity score adjusts for some between-person confounding, there is still residual confounding leading to biased estimators. This residual confounding may be a combination of both between-person confounding that the method did not fully address or unmeasured confounding that the method is not able to consider, as well as within-person confounding that the design is not intended to account for [13].

Using our simulated dataset, we evaluated how often the 95 % confidence intervals contained the true effect size. We observed that the coverage was far below the nominal 95 % value, with estimated intervals covering less than 55 % of true effects under study for all outcomes. This indicates that large variability of the estimates is not fully captured in the confidence intervals produced by the analysis.

Our findings suggest the cohort method can provide supporting evidence to evaluate an effect of a medical product, but should not be considered definitive evidence given the degree of error observed within the effect estimates. Confidence intervals should be interpreted with caution, unless they can be demonstrated to be properly calibrated. These results are consistent with the findings from prior experiments conducted across 53 drug-outcome pairs within 10 disparate observational databases, which found that a propensity-score adjusted new user cohort design was reasonably predictive (AUC = 0.68–0.84) in discriminating positive controls from negative controls, but that traditional interpretations of statistical significance would result in an unacceptably high false positive rates [45]. Further research is required to confirm these results and to evaluate their generalizability across other outcomes and other databases. While this experiment assessed many different configurations of the new user cohort design, many more still exist and some unexplored combinations may yield improved performance from what has been observed to date. In particular, we focused the evaluation on the estimated odds ratios from a logistic regression, but the cohort design can also be used to estimate risk differences, rate ratios through Poisson regression or hazard ratios from Cox proportional hazards models. Furthermore, while many of the initial experiments centered on decisions associated with the design and application of the propensity score, we found that comparator selection had a major impact on the method's overall performance. While the Brookhart and Bayesian regression approaches to covariate selection may theoretically risk identifying instrumental variables that could bias the propensity score, our real-world results are consistent with simulation findings from Myers et al. [46]

which show that, in practice, this potential bias is small. Additional experiments that expand the universe of test cases and further explore the interplay between database, exposure, outcome and analysis choices could shed light on best practices for developing prospective cohort studies for signal generation, refinement, and evaluation moving forward.

Our results suggest that special consideration is required to identify and evaluate an appropriate comparator group, keeping in mind that the choice should reflect the scientific question of interest. We evaluated five different comparator selection approaches, but each has its own limitations. When selecting a single drug that shares the same indication, we have selected the most prevalent drug in a different therapeutic class, but the most prevalent drug may not necessarily be the most comparable to the target drug nor the alternative which is least likely to be itself associated with the outcome. Combining all drugs with the same indication introduces potential heterogeneity in the comparator group and may not be recommended unless there was sufficient confidence that each product within the set were sufficient similar. Creating a non-user cohort of patients with the indication presents well-defined challenges with immortal time bias [11]. The use of 'negative control' drugs or the entire database as a comparator group is likely to introduce selection bias, since patients without the indication likely have differential prognosis regarding certain outcomes and may include the 'healthy user' population which could positively bias any effect estimates for any treatment for any disease.

In the context of the drug safety question of whether a product increases a risk, the comparator group is serving as a proxy for the counterfactual of 'exposed patients having not been exposed'. To achieve this objective, the comparator group should ideally represent patients that are similar in all regards to those in the exposed population at the time of treatment initiation, but capture a time-at-risk that is not associated with the outcome of interest. In this regard, the comparator group is intended to serve as a negative control [47]. Defining a comparator group that definitively satisfies both criteria can be difficult: the patients that are most like those exposed to the target drug are generally those patients diagnosed with the same indication and initiating an alternative treatment.

However, the true effect of the alternative treatment on the outcome is likely unknown, or at least uncertain, if it is in any way related pharmacologically to the target drug of interest. For example, if we want to study an adverse event with an unknown association with lisinopril, then a comparison with an angiotensin receptor blocker may be advantageous from the perspective of minimizing confounding by indication [48] but disadvantageous from the perspective of having confidence that the adverse event isn't also associated with the comparator. Without external information, evaluating this 'no effect' criteria using the comparative cohort design presents a circular conundrum: in order to determine that a comparator is not associated with an outcome, it must be compared against some other comparator known not to be associated with the outcome, which in turn must be compared ad infinitum.

Conversely, selecting a benign drug as an active comparator may more easily satisfy the condition that it not be associated with the outcome of interest, but likely presents a challenge in assuring that patients are truly comparable to those exposed to the target drug. The strategy of using a 'negative control' drug as an active comparator group has been discussed previously [9]. Our results suggest this strategy may offer advantages to using an active comparator from the same indication. One possible explanation for this finding is that, of the two criteria required to be an appropriate comparator, it is easier to adjust for confounding due to imbalances in populations than it is to address the latent causal effects of the comparator on the outcome. When evaluating a method's ability to identify risk, (ie. to distinguish effect from no effect) it may be useful to have the comparator group for each drug be more consistent with each other by using the same comparator group. Any bias present in the comparator group will be the same across drugs, and not effect the relative ranking of drugs in the evaluation or a possible risk identification system. Of course, when directly comparing the relative safety of two drugs, one might do without the negative control comparator, although we have not evaluated the performance of such direct comparisons.

Our study has several limitations that should be considered when interpreting these results. Even though the ground truth set was meticulously constructed, there is never unequivocal evidence of the true status of a drug-outcome pair. It is certainly possible that some of our negative controls may prove to be positive in the future, and vice versa. Furthermore, in order to test the design on a large scale, we had to automate the design decisions that are normally made by experts. In particular situations, such experts might have adjusted the design to meet unique circumstances, for instance by adding additional exclusion criteria for subjects. Experts may consider different comparators for different outcomes, or may make a custom selection outside of the scope for the five strategies studied here. Similarly, experts may argue that they would opt to custom tailor the time-at-risk window for every unique drug-outcome scenario, which would be inconsistent with our standardized use of a constant time-at-risk definition across all test cases within a given outcome. Stang et al. have shown that there is considerable heterogeneity among analysis choices that the community may make when faced with similar research questions [49]. Although it is impossible to say whether such ad hoc changes lead to an improvement or deterioration of the performance, it does mean our results cannot easily be generalized to all new user cohort studies. Another potential limitation is that positive and negative controls may differ in other respects as well. For instance, in general, positive control drugs tend to be used longer than negative control drugs, which could bias the results. In this work, we also made no attempt to control for time-dependent confounding, and analyses ignored differences in dosage and cumulative drug effects which may exist in real data but were not explicitly modeled in simulation. While we excluded patients with overlapping time-at-risk in both target and comparator cohorts, further refinements could be considered to preserve the non-overlapping portions and model the time-dependent transition between treatments. Given that these factors may be important for studying effects of chronic drug exposure, marginal structural modeling could be applied within a new user cohort framework as a possible future direction. Additionally, while we have applied the new user cohort design in a manner consistent with the recommendations of Ray [8] and Schneeweiss [12] and the implementations within the Sentinel Initiative [40–44], it is important to highlight that other approaches can be considered. Rather than restricting only to patients who start therapy for the target medication, it could be useful to further exclude patients who have any past exposure to any therapies which may have been given for the indication of the target medication. This approach would be limited to use for evaluation of first-line therapies, but could provide additional confounding control if number of prior treatments for the indication was regarded as a potential source of bias. Further work is also needed to evaluate methods for comparing newly marketed products with existing therapies, where the dynamics of the new user cohort design and the adjustment strategy [50] using propensity scores and/or disease risk scores may need to be tailored, to achieve improved predictive accuracy.

A key challenge the community faces in the application and interpretation of the cohort design for drug safety is drawing the distinction between identifying the risks of a specific treatment and comparing the relative effectiveness of alternative treatments. The questions “Does the drug cause the adverse event?” and “Does the drug have a higher risk of the adverse event than an alternative treatment?” are both valid and important, but may require different analytical approaches, proper interpretation, and careful communication of the results. It is all too easy for a comparative statement of relative effects to be misconstrued into an absolute statement of the safety of a product. Just as it may not be appropriate to declare a safety signal on the basis of a comparative imbalance between alternative treatments, it may be premature to declare a drug ‘safe’ if a comparative analysis shows no difference. In order for a comparative assessment to lend evidentiary support to a causal safety question, the comparator group should be evaluated to assure the exchangeability assumption can be adequately upheld and independently assessed to ensure it does not itself carry a causal effect of the outcome. Developing a standardized approach for evaluating the adequacy of a comparator could be a valuable contribution to the field. Until then, given the complexities of satisfying these criteria, multiple comparators may be considered to be used as part of a comprehensive sensitivity analysis, alongside of the myriad of design decisions commonly explored within each comparison.

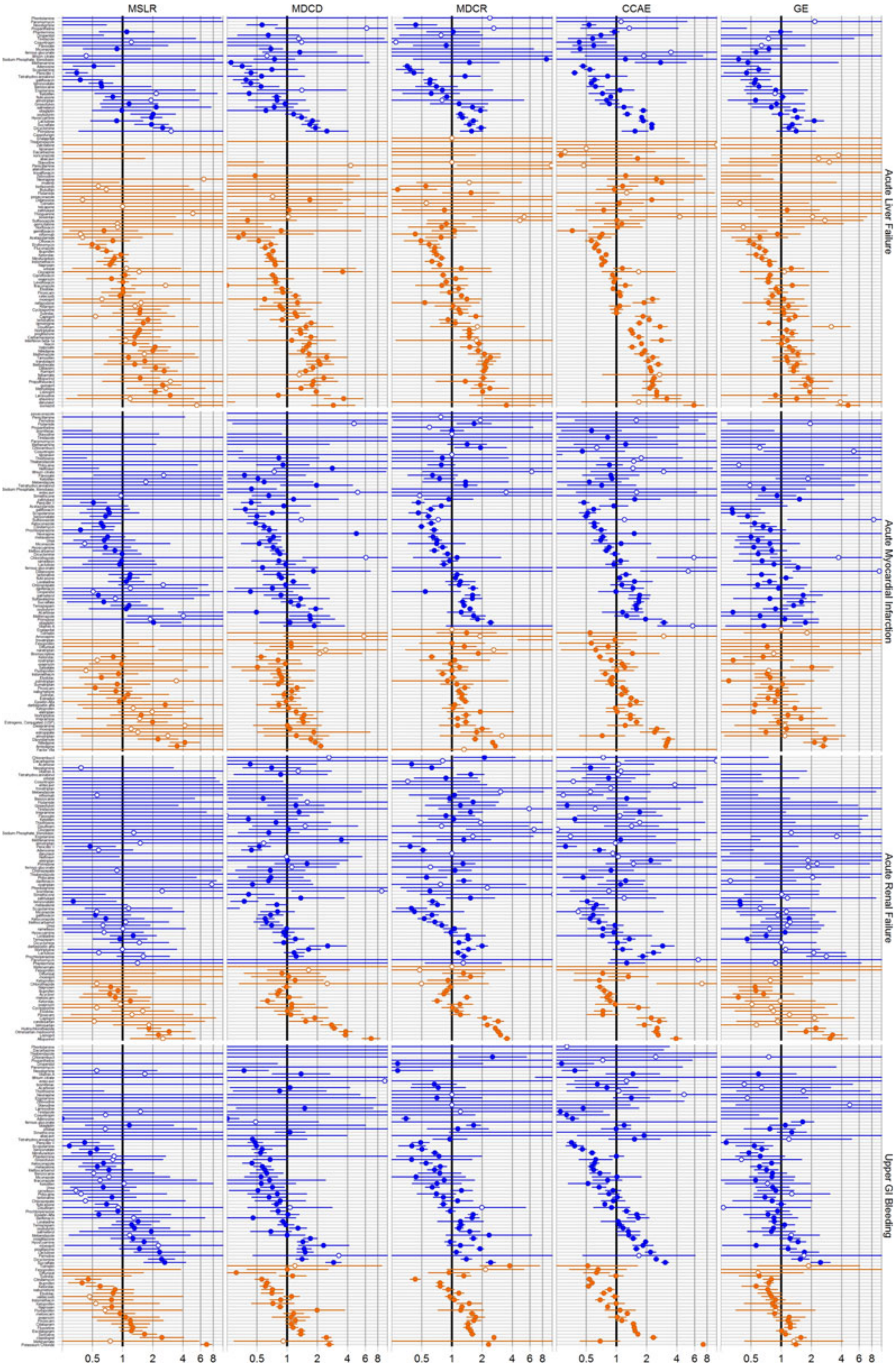
5 Conclusions

The new user cohort method can contribute useful information toward a risk identification system, but careful consideration of the analysis choices and proper interpretation of the study findings in the context of the method’s operating characteristics are required to support its use. The modest predictive accuracy, substantial variance in the error distribution, and low coverage probability observed in experiments from both real data and simulation suggest that additional caution is needed when evaluating cohort results for drug safety. Specific attention is required to select the comparator group and evaluate its appropriateness for use in either assessing the causal effect of a specific product or examining the comparative effectiveness of alternative treatments.

Acknowledgements 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 previously received funding from FNIH. Susan Gruber and Ivan Zorych have no conflicts of interest to declare.

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 honoraria 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.

Appendix



Incident user cohort design estimates for all test cases, by database. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity. *Blue* negative controls; *Orange* positive controls; *each line* represents point estimate and 95 % confidence interval for the drug-outcome pair in a particular database

References

1. FDA. Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment. US FDA Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research; 2005.
2. Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol. 2002;31(2):422–9.
3. Hofler M. Causal inference based on counterfactuals. BMC Med Res Methodol. 2005;5:28.
4. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health. 2006;60(7):578–86.
5. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413–9.
6. FDA. The sentinel initiative: a national strategy for monitoring medical product safety. May 2008 (cited 2012 September 15). Available from: http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm089474.htm.
7. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Ann Rev Public Health. 2000;21:121–45.
8. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol. 2003;158(9):915–20.
9. Schneeweiss S, Patrick AR, Sturmer T, Brookhart MA, Avorn J, Maclure M, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. Med Care. 2007;45(10 Supl 2):S131–42.
10. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology. 2009;20(4):488–95.
11. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol. 2008;167(4):492–9.
12. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010;19(8):858–68.
13. 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.
14. Johnson ES, Bartman BA, Briesacher BA, Fleming NS, Gerhard T, Kornegay CJ, et al. The incident user design in comparative effectiveness research. Pharmacoepidemiol Drug Saf. 2013;22(1):1–6.
15. van Staa TP, Abenhaim L, Leufkens H. A study of the effects of exposure misclassification due to the time-window design in pharmacoepidemiologic studies. J Cli Epidemiol. 1994; 47(2):183–9.
16. McMahon AD, Evans JM, McGilchrist MM, McDevitt DG, MacDonald TM. Drug exposure risk windows and unexposed comparator groups for cohort studies in pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 1998;7(4):275–80.
17. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009;20(4):512–22.
18. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. Am J Epidemiol. 2011;173(12):1404–13.
19. Wahl PM, Gagne JJ, Wasser TE, Eisenberg DF, Rodgers JK, Daniel GW, et al. Early steps in the development of a claims-based targeted healthcare safety monitoring system and application to three empirical examples. Drug Saf Int J Med Toxicol Drug Experience. 2012;35(5):407–16.
20. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149–56.
21. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol. 2006;163(3):262–70.
22. Sturmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution–a simulation study. Am J Epidemiol. 2010;172(7):843–54.
23. Ryan PB, Schuemie MJ. Evaluating performance of risk identification methods through a large-scale simulation of observational data. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0110-2
24. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A comparison of the empirical performance of methods for a risk identification system. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0108-9
25. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54–60.
26. 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.
27. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
28. Brookhart MA. Incident User Design (IUD-HOI) 2010 (cited 2013 January 28). Available from: http://omop.org/MethodsLibrary.
29. Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. Technometrics. 2007;49:291–304.
30. Observational Medical Outcomes Partnership June 2012 Symposium presentations 2012 (cited 2013 January 23). Available from: http://omop.org/2012SymposiumPresentations.
31. 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. 2013 (in this supplement issue). doi:10.1007/s40264-013-0097-8
32. 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.
33. Cantor SB, Kattan MW. Determining the area under the ROC curve for a binary diagnostic test. Med Decis Mak Int J Soc Med Decis Mak. 2000;20(4):468–70.
34. Ebell MH, Smith MA, Barry HC, Ives K, Carey M. The rational clinical examination. Does this patient have strep throat? J Am Med Assoc. 2000;284(22):2912–8.
35. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. New Engl J Med. 2005;353(17):1773–83.
36. Martin BJ, Finlay JA, Sterling K, Ward M, Lifsey D, Mercante D, et al. Early detection of prostate cancer in African-American men through use of multiple biomarkers: human kallikrein 2 (hK2), prostate-specific antigen (PSA), and free PSA (fPSA). Prostate Cancer Prostatic Dis. 2004;7(2):132–7.
37. Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol. 1993;46(10):1075–9; discussion 81–90.
38. Tisdale J, Miller D. Drug-induced diseases: prevention, detection, and management. 2nd ed. American Society of Health-System Pharmacists, Bethesda; 2010.

39. 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.
40. FDA Drug Safety Communication: Update on the risk for serious bleeding events with the anticoagulant Pradaxa (dabigatran). November 2, 2012 (cited 2012 December 1). Available from: http://www.fda.gov/Drugs/DrugSafety/ucm326580.htm.
41. Platt R. Mini-Sentinel Program to evaluate the safety of marketed medical products—progress and direction. ISPE—a symposium at the 27th international conference on pharmacoepidemiology. Chicago, IL; 2011.
42. Fireman B, Toh S, Butler MG, Go AS, Joffe HV, Graham DJ, et al. A protocol for active surveillance of acute myocardial infarction in association with the use of a new antidiabetic pharmaceutical agent. Pharmacoepidemiol Drug Saf. 2012;21(Suppl 1):282–90.
43. Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. Pharmacoepidemiol Drug Saf. 2012;21(Suppl 1):41–9.
44. Robb MA, Racoosin JA, Worrall C, Chapman S, Coster T, Cunningham FE. Active surveillance of postmarket medical product safety in the Federal Partners' Collaboration. Med Care. 2012;50(11):948–53.
45. 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.
46. 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.
47. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383–8.
48. Walker AM. Confounding by indication. Epidemiology. 1996;7(4):335–6.
49. Stang PE, Ryan PB, Overhage JM, Schuemie MJ, Hartzema AG, Welebob E. Variation in choice of study design: findings from the Epidemiology Design Decision Inventory and Evaluation (EDDIE) Survey. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0103-1
50. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. Pharmacoepidemiol Drug Saf. 2012;21(Suppl 2):138–47.