Drug Saf (2013) 36 (Suppl 1):S159–S169
DOI 10.1007/s40264-013-0109-8

ORIGINAL RESEARCH ARTICLE

Replication of the OMOP Experiment in Europe: Evaluating Methods for Risk Identification in Electronic Health Record Databases

Martijn J. Schuemie · Rosa Gini · Preciosa M. Coloma · Huub Straatman · Ron M. C. Herings · Lars Pedersen · Francesco Innocenti · Giampiero Mazzaglia · Gino Picelli · Johan van der Lei · Miriam C. J. M. Sturkenboom

Abstract

Background The Observational Medical Outcomes Partnership (OMOP) has just completed a large scale empirical evaluation of statistical methods and analysis choices for risks identification in longitudinal observational healthcare data. This experiment drew data from four large US health insurance claims databases and one US electronic health record (EHR) database, but it is unclear to what extend the findings of this study apply to other data sources.

Objective To replicate the OMOP experiment in six European EHR databases.

Research Design Six databases of the EU-ADR (Exploring and Understanding Adverse Drug Reactions) database network participated in this study: Aarhus (Denmark), ARS (Italy), HealthSearch (Italy), IPCI (the Netherlands), Pedianet (Italy), and Pharmo (the Netherlands). All methods in the OMOP experiment were applied to a collection of 165 positive and 234 negative control drug–outcome pairs across four outcomes: acute liver injury, acute myocardial infarction, acute kidney injury, and upper gastrointestinal bleeding. Area under the receiver operator characteristics curve (AUC) was computed per database and for a combination of all six databases using meta-analysis for random effects. We provide expected values of estimation error as well, based on negative controls.

Results Similarly to the US experiment, high predictive accuracy was found (AUC >0.8) for some analyses. Self-controlled designs, such as self-controlled case series, IC

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.

Electronic supplementary material The online version of this article (doi:10.1007/s40264-013-0109-8) contains supplementary material, which is available to authorized users.

M. J. Schuemie (✉) · R. Gini · P. M. Coloma ·
R. M. C. Herings · J. van der Lei · M. C. J. M. Sturkenboom
Department of Medical Informatics, Erasmus University
Medical Center, P.O. Box 2040, 3000 CA Rotterdam,
The Netherlands
e-mail: m.schuemie@erasmusmc.nl

R. Gini · F. Innocenti
Agenzia Regionale di Sanità della Toscana, Florence, Italy

H. Straatman · R. M. C. Herings
PHARMO Institute, Utrecht, The Netherlands

L. Pedersen
Department of Clinical Epidemiology, Aarhus University
Hospital, Århus, Denmark

F. Innocenti · G. Mazzaglia
Health Search, Italian College of General Practitioners, Florence,
Italy

G. Picelli
Pedianet, Società Servizi Telematici SRL, Padova, Italy

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

temporal pattern discovery and self-controlled cohort achieved higher performance than other methods, both in terms of predictive accuracy and observed bias.

Conclusions The major findings of the recent OMOP experiment were also observed in the European databases.

1 Introduction

There has been an increased interest in using longitudinal observational healthcare data for identification of risk of adverse events associated with prescription drugs. In the past this type of data has been used extensively to investigate a single drug and a single health outcome of interest (HOI) at a time, but more recently there has been a call to use these data for active safety surveillance, studying many drugs simultaneously, in complement to spontaneous reporting systems such as the FDA's AERS database. Several studies have investigated the performance of a wide array of statistical risk identification methods that can be applied to this data, showing promising results [1, 2]. The Observational Medical Outcomes Partnership (OMOP) recently completed a large empirical study evaluating more methods and analysis choices using a larger reference set than in their first evaluation [3], with the goal of determining which analyses are best suited for specific circumstances, and what the operating characteristics of these methods are in those circumstances. The results showed that, within the databases included in the study, several analyses achieved high predictive accuracy in distinguishing positive from negative control drug–outcome pairs, but that the estimates of all methods were biased and the 95% confidence interval rarely contained the true relative risk in those cases where we know the true relative risk.

These findings from OMOP can inform us which method to choose, which analysis options to pick, which type of database(s) to query and how to interpret the observed relative risk estimates when studying drugs in relation to one of the four health outcomes of interest (HOIs) defined in the study: acute liver injury, acute myocardial infarction, acute kidney injury, and upper gastrointestinal bleeding. These findings have a heuristic value as soon as the following assumption of exchangeability is considered to be reasonable: if we were to screen a new drug to detect whether it has one of the four HOIs as an adverse reaction, the probability of correctly detecting existing, and of correctly rejecting non-existing reactions (with a given method and in a given database) is similar, respectively, to the corresponding probabilities among the positive and negative control pairs in this experiment. We consider this assumption reasonable if we are only concerned with adverse reactions that take place in a similar timeframe as the one reflected in the positive control sample (i.e. short-time acute adverse reactions). Measuring similar findings to OMOP, but in different databases, should support this assumption of applicability.

The most recent OMOP experiment drew data from four large US health insurance claims databases and one US electronic health record (EHR) database: MarketScan® Lab Supplemental (1.2 m persons), MarketScan® Medicare Supplemental Beneficiaries (4.6 m persons), MarketScan® Multi-State Medicaid (10.8 m persons), MarketScan® Commercial Claims and Encounters (46.5 m persons), and the GE Centricity™ (11.2 m persons) EHR database. We wished to test whether the findings of the study were in fact attributable to unobserved characteristics that are unique to healthcare data collected in the US. We have therefore replicated the experiment in six databases of the European EU-ADR (Exploring and Understand Adverse Drug Reactions) network: Aarhus (Denmark), ARS (Italy), HealthSearch (Italy), IPCI (the Netherlands), Pedianet (Italy), and Pharmo (the Netherlands). These databases are very different from the US databases, include EHR and (non-commercial) administrative claims data, and contain data from healthcare systems that capture data about the general population, although usually restricted to those living in a specific region of a country. For instance, the Aarhus database captures all information about the population, but only in the Aarhus region. In healthcare systems which do not provide universal access to healthcare, a database population is selected according to eligibility criteria such as socio-economic status, which may impact the distribution of risk factors and morbidity/mortality. This is the case for example in OMOP databases related to the Medicare or Medicaid programs.

All seven methods included in the most recent OMOP experiment were applied in each database to a collection of 165 positive control and 234 negative control drug–outcome pairs across all four HOIs. Each method required specification of several analysis choices, such as the time-at-risk window definition, number of controls per case, or which variables to include in the propensity score, and a large number of possible analysis choices were explored. For each analysis the predictive accuracy was computed as the area under the receiver operator characteristics curve (AUC). We provide expected values of estimation error as well, based on negative controls.

2 Methods

Table 1 shows an overview of the contributing databases. These are the same databases that also participated in an earlier comparison of statistical methods [2], although one of the databases in that study (UNIMIB) could not participate in this experiment. The data extraction process used in that earlier study was reused, resulting in information on all drug dispensing or prescribing, information on patients

Table 1 Characteristics of the participating databases

	Aarhus Denmark	ARS Italy	Health Search Italy	IPCI Netherlands	Pedianet Italy	PHARMO Netherlands
Source population	2,000,000	4,000,000	1,000,000	750,000	140,000 children	1,280,000
Type of database	National health registry with:	Record linkage system with:	General practice database (no children)	General practice database	General practice pediatric database	Record linkage system with:
	(1) Registry inhabitants	(1) Registry inhabitants				(1) Registry inhabitants
	(2) Regional Drug dispensation records	(2) Regional Drug dispensation records				(2) Regional Drug dispensation records
	(3) Hospital claims database	(3) Hospital claims database				(3) Hospital claims database
	(4) Lab values	(4) Death registry				(4) Lab values
	(5) Death registry					
Diagnosis coding scheme	ICD-10	ICD-9CM	ICD-9CM	ICPC	ICD-9CM	ICD-9CM
Language of free text	No free text	No free text	Italian	Dutch	Italian	No free text

Table 2 Crude incidence rates (per 100,000 patient years) for the four health outcomes of interest in the five OMOP databases and the six EU-ADR databases

	OMOP					EU-ADR					
	MSLR	MDCD	MDCR	CCAE	GE	Aarhus	ARS	Health search	IPCI	Pedianet	PHARMO
Acute liver failure	2,870	1,325	1,980	1,266	833	84	11	49	547	27	5
Acute renal failure	335	472	1,512	117	56	28	34	6	415	1	4
Acute myocardial infarction	1,797	1,071	4,988	750	626	202	208	96	777	0	144
Upper GI bleeding	1,488	998	2,571	802	486	88	64	93	526	101	43

(gender, date-of-birth, start and end of follow up), and information on the occurrences of HOIs.

In EU-ADR, drug prescriptions are encoded using the ATC (Anatomic Therapeutic Chemical) system, which were translated for this experiment to RxNorm identifiers using the OMOP Standard Vocabulary v2.1 [4]. This mapping was manually inspected, and when no matching RxNorm term was found automatically this was added manually. The complete ATC to RxNorm mapping table is provided in ESM.

In the OMOP experiment the HOIs were uniformly extracted across databases using the same algorithm. In EU-ADR, the HOIs were extracted using definitions that were harmonized across databases using an iterative procedure described elsewhere [5]. Because of this, each database used its own extraction algorithm. Since the HOIs in this experiment are a subset of those used in earlier experiments, no new definitions needed to be created. Table 2 shows the crude incidence rates for the four HOIs in the EU-ADR databases and the OMOP databases. The lower incidence rates in EU-ADR can be attributed to more specific outcome definitions than those used in OMOP. For example, in OMOP the acute liver failure definition included ICD code 277.4 ‘Disorders of bilirubin excretion’, whilst the EU-ADR definition did not include this code or any of its equivalents in other coding systems. In the EU-ADR databases containing information on hospitalizations (Aarhus, ARS and Pharmo), secondary diagnoses in hospitalization were ignored for the four outcomes reported here, whilst they were included in OMOP together with diagnoses during specialist or primary care visits. The higher incidences in IPCI are most likely due to the free-text queries that were used which have lower specificity, and Pedianet is restricted to children, who have lower incidence rates than adults. The MarketScan$^{®}$ Medicare Supplemental Beneficiaries database contains only older subjects, who have higher incidence rates than the general population. Other possible reasons for the differences could be under-registration in the EU-ADR databases, or over-registration in the US insurance claims databases as a result of coding behavior seeking the highest imbursements [6]. We have added an age-stratified analysis in Appendix 2 in ESM showing that some but not all differences are explained by differences in age distribution.

Some of the analysis choices of the case-control method required visit dates as an indication of health care utilization. This was approximated by assuming every date of prescribing, dispensing or occurrence of an HOI indicated a visit. Seven methods (i.e., overall study designs) were evaluated:

- **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)** are a suite of methods borrowed from data-mining in spontaneous reports, including proportional reporting ratios (PRR), reporting odds ratios (ROR), 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 design (SCC)** 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 conditioned on the person [13].

For each method, a number of different analysis choices could be made. Each analysis (i.e. combination of analysis choices) included in the experiment was given a unique identifier. For example, SCC: 403002 is the unique identifier that reflects the analysis for SCC which uses all occurrences of drug exposure and all occurrences of the outcome, defined both the time-at-risk and the control period as the length of exposure +30 days, and classified events occurring on the index exposure date as within the time-at-risk.

The same analyses were evaluated as in the most recent OMOP experiment [1], except for those with specific settings requiring information on conditions other than the HOIs, for instance conditions that are known indications for drugs. Because the databases use widely different ways to record conditions, including free text in two languages and ICD-9, ICD-10 and ICPC codes, extraction of large numbers of conditions in a uniform way was not feasible in the scope of this experiment. These analysis choices were therefore not evaluated in this experiment.

The methods were executed using all analysis choice combinations against 399 drug–outcome pairs to generate an effect estimate and standard error for each pair and analysis. These test cases include 165 ‘positive controls’—active ingredients with evidence to support a positive causal association with the outcome—and 234 ‘negative controls’—active ingredients with no evidence to expect a causal effect on the outcome, and were limited to four HOIs: 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 [14]. 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 [15]. We also combined the estimates from all databases using meta-analysis for random effects, and computed separate performance statistics for this combination.

3 Metrics

To gain insight into the ability of an analysis to distinguish between positive and negative controls the relative risk estimates were used to draw a receiver operator characteristics (ROC) curve, by indicating the sensitivity and the complement to specificity (i.e., 1 − specificity) for all choices of threshold for the effect: if for instance the point (0.6,0.2) belongs to the curve this means that when the corresponding threshold is chosen, 60 % of positive pairs are correctly classified as adverse reactions but 20 % of negative pairs are inaccurately classified as adverse reactions as well. The area under the ROC curve (AUC) was computed, a measure of predictive accuracy [16]: an AUC of 1 indicates a perfect prediction of which test cases are positive, and which are not. An AUC of 0.5 is equivalent to random guessing.

Often we are not only interested in whether there is an effect or not, but would also like to know the magnitude of the effect. However, in order to evaluate whether a method produces correct relative risk estimates, we must know the true effect size. In real data, this true effect size is never known with great accuracy for positive controls, and we must restrict our analysis to the negative controls, where we assume that the true relative risk is 1. Elsewhere, simulation studies were used to evaluate the accuracy of the estimates when the true relative risk is larger than 1 [17].

4 Results

Figure 1 shows the number of positive and negative controls for which there was enough power to detect a relative risk of 1.25 in each database. In Pedianet, the smallest database, there were no drug–outcome pairs that met this criterion, so database-specific performance could not be computed. However, Pedianet was included in the meta-analysis. For several other databases we also see low numbers of drug–outcome pairs for which there was enough power, especially for acute liver failure and acute renal failure because both these outcomes are rare.

In Fig. 2, each dot (both light and dark) represents the AUC of a particular analysis on a particular database or the meta-analysis. To avoid unstable estimates, the AUC was not computed when fewer than five positive or negative controls were available. As can be seen, each method shows a wide variation in performance with different settings, and the best performing settings per database are detailed in Table 3. The analyses with the highest AUC for the meta-analysis are specified in Table 4.

For each analysis we computed the average AUC across all HOIs and databases (not including the meta-analysis) where an AUC could be computed, and determined those analysis with the highest average AUC. These are plotted in Fig. 2 using the lines and darker colors, and the details and average AUC are provided in Table 5.

When comparing these average AUCs, we see that self-controlled designs such as SCC, ICTPD and SCCS in general tend to perform better than others. If we compare the average AUCs of these analyses in both the US OMOP experiment and this replication, as shown on the right side in Table 6, we see similar performance despite the large differences in underlying databases. Similarly, if we use the average overall best performing analyses found in OMOP and compute the average AUCs in EU-ADR, we see that these settings achieve almost exactly the same performance as the EU-ADR defined optimal settings. For example, the best performing analysis in OMOP was SCC: 403002, which achieved an average AUC of 0.72 in EU-ADR data, compared with an AUC of 0.82 in the OMOP data. The best performing analysis in EU-ADR is SCC: 409013, achieving an AUC of 0.75 in EU-ADR data, and

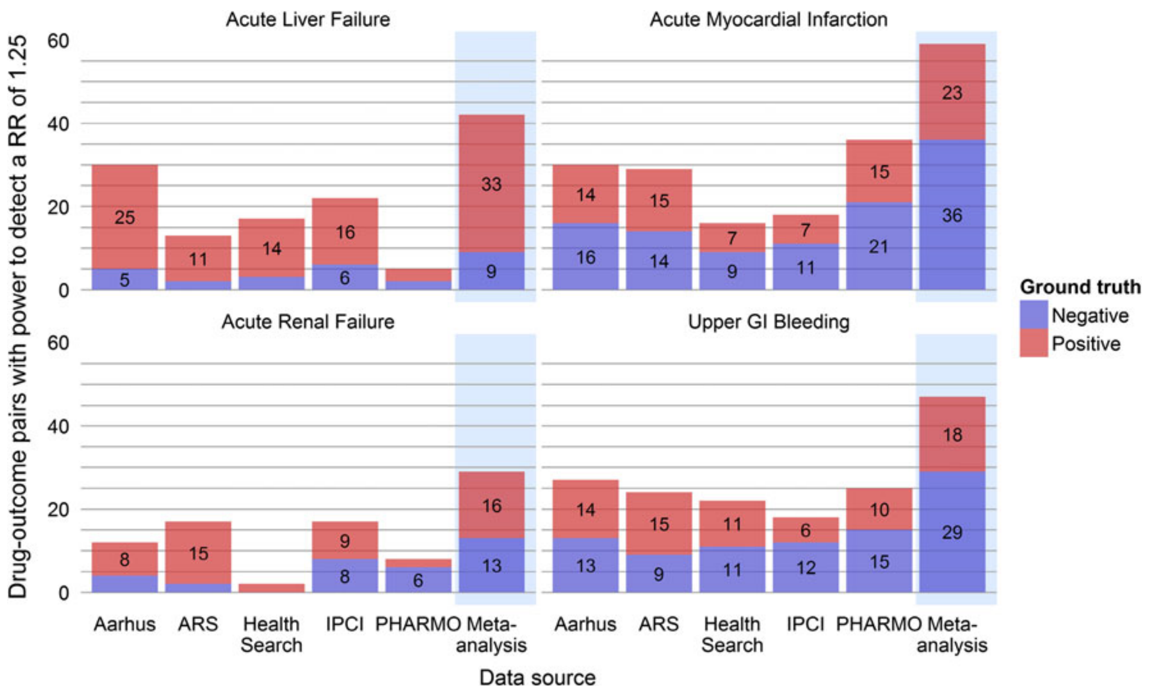

Fig. 1 Numbers of positive and negative control drug–outcome combinations for which there is enough power to detect a relative risk of 1.25, per outcome of interest and database and meta-analysis

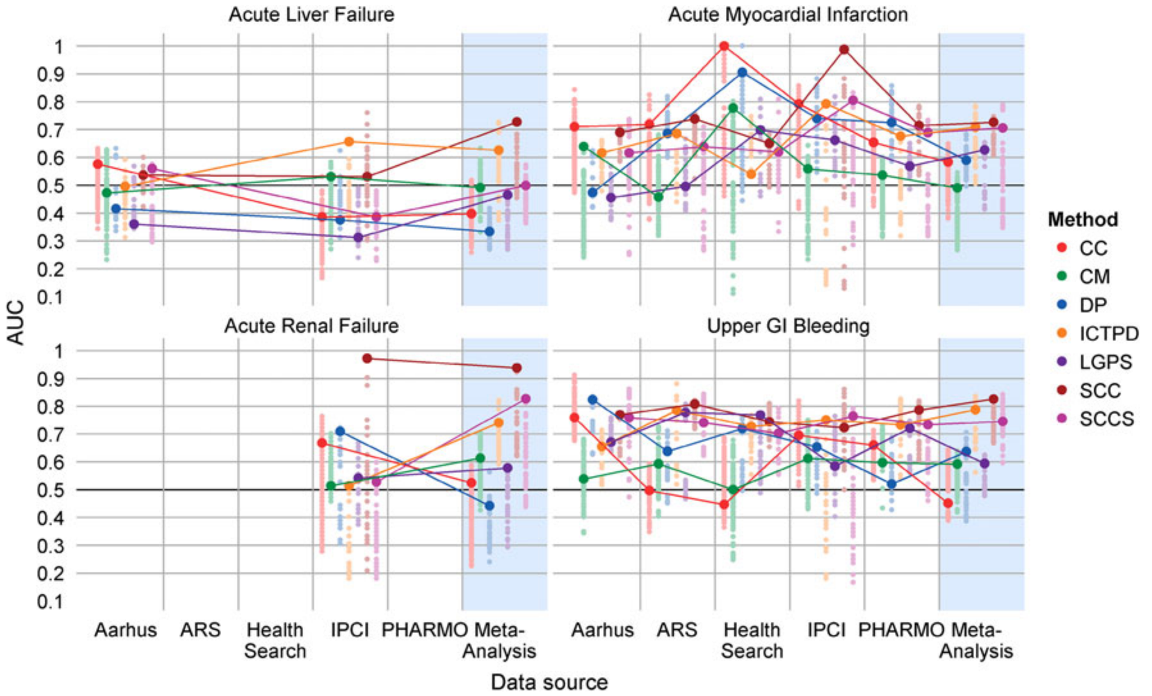

Fig. 2 Area under the receiver operator characteristics curve (AUC). *Light dots* indicate the performance of all different analyses. *Dark colors* and *lines* indicate the analyses that achieved the highest AUC averaged over all outcomes and databases with enough data. (*CC* case–control, *CM* cohort method, *DP* disproportionality methods, *ICTPD* Information Component Temporal Pattern Discovery, *LGPS* Longitudinal Gamma Poisson Shrinker, *SCC* Self-Controlled Cohort design, *SCCS* self-controlled case series). (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)

an AUC of 0.80 in OMOP data. What is also striking is that both in OMOP and EU-ADR the same three methods comprise the top three: SCC, ICTPD and SCCS.

Using the analysis with the highest average AUC as representative analysis for each method, Fig. 3 shows the estimates for the negative control drugs for which there was enough power. Because these drugs are believed not to cause the outcome, we assume the true relative risk is one. We see that most methods produce estimates that are very different from one, with most values between 0.5 and 2, with a tendency for ICTPD, LGPS, SCC and SCCS to be positively biased.

5 Discussion

In this replication experiment a large array of analyses was evaluated using data of over 9 million persons in Europe. Several analyses achieved high AUCs both in individual databases and in a meta-analysis. Although for meta-analysis higher AUCs were not observed, more drug–outcome pairs were statistically powered and could therefore be studied.

Based on the results, several recommendations can be made for selecting the appropriate method and analysis choices, and interpreting the results when studying the next drug–outcome pair. In the recommendations we focus on AUC, since a first test that must be passed by a method is whether it can distinguish between an effect and no effect. Only after that should we consider whether the observed effect is an approximation of the true effect size.

With respect to selecting the method, data from both experiments clearly indicate that self-controlled designs perform better than other designs. Especially SCC has proven robustly to outperform all other methods across most outcomes and data sources. A possible explanation for these findings is that in observational data there is potential for serious confounding, and given the lack of variables that are typically captured in these databases, only a fraction of this confounding can be dealt with, for instance through propensity scores. Self-controlled methods are at least immune to between-person confounding, and are therefore better able to identify potential causal relations. Arguably, the remaining within-person bias is less associated with a specific drug, and is more general: the time of drug exposure is correlated with reduced overall health and increased healthcare utilization, and is therefore more likely to show any outcomes, including the HOIs. We therefore suspect that all self-controlled designs are positively biased, but that this bias is present in both positive and negative controls.

When considering the best settings for the method of choice, the data show that the average overall optimal

Table 3 Best performing analyses as measured by AUC, per database and outcome of interest

	Acute liver failure	Acute myocardial infarction	Acute renal failure	Upper GI bleeding
Aarhus	AUC = .63 (CC:2000164) - Up to 100 controls per case - Require 30 days observation time - Risk period: all time post exposure - Don't include index date - Match on age, sex, and visit (180 days) - No nesting - Include all exposures	AUC = .84 (CC:2000213) - Up to 100 controls per case - Require 180 days observation time - Risk period: 30 days from exposure start - Don't include index date - Match on age, sex, and visit (30 days) - No nesting - Include first exposure only	Not enough data to compute AUC	AUC = .91 (CC:2000213) - Up to 100 controls per case - Require 180 days observation time - Risk period: 30 days from exposure start - Don't include index date - Match on age, sex, and visit (30 days) - No nesting - Include first exposure only
ARS	Not enough data to compute AUC	AUC = .82 (CC:2000349) - Up to 10 controls per case - Require 180 days observation time - Risk period: all time post exposure - Don't include index date - Match on age, sex, and visit (30 days) - No nesting - Include first exposure only	Not enough data to compute AUC	AUC = .88 (ICTPD:3064001) - Control period: 180 days prior exposure - Risk period: 60 days from exposure start - Don't use 1 day before exposure - Use 1 m prior in expected calculation
Healthsearch	Not enough data to compute AUC	AUC = 1.00 (CC:2000157) - Up to 10 controls per case - Require 30 days observation time - Risk period: all time post exposure - Don't include index date - Match on age, sex, and visit (30 days) - No nesting - Include first exposure only	Not enough data to compute AUC	AUC = .81 (SCC:402001) - Include all exposures - Include all outcome occurrences - Risk period: 30 days from exposure start - Include index date in risk period - Control period: 30 days prior exposure - Include index date in control period
IPCI	AUC = .76 (SCC:407002) - Include all exposures - Include first outcome only - Risk period: length of exposure +30 days - Don't include index date in risk period - Control period: length of exposure +30 days - Don't include index date in control period	AUC = .99 (SCC:409013) - Include first exposure only - Include first outcome only - Risk period: all time post exposure - Don't include index date in risk per. - Control period: all time prior - Don't include index date in control period	AUC = .97 (SCC:409013) - Include first exposure only - Include first outcome only - Risk period: all time post exposure - Don't include index date in risk period - Control period: all time prior - Don't include index date in control period	AUC = .86 (SCC:405002) - Include first exposure only - Include all outcome occurrences - Risk period: length of exposure +30 days - Don't include index date in risk period - Control period: length of exposure +30 days - Don't include index date in control period

Table 3 continued

	Acute liver failure	Acute myocardial infarction	Acute renal failure	Upper GI bleeding
Pharmo	Not enough data to compute AUC	AUC = .86 (DP:106003) - Include all outcome occurrences - Stratify by age and sex - Risk window: length of exposure +60 days - Use multi-item GPS	Not enough data to compute AUC	AUC = .83 (ICTPD:3052001) - Control period: 180 days prior exposure - Risk period: 60 days from exposure start - Don't use 1 day before exposure - Don't use 1 m prior in expected calculation

AUC area under the receiver operator characteristics curve, *CC* case control, *DP* disproportionality methods, *ICTPD* Information Component Temporal Pattern Discovery, *SCC* self controlled cohort

Table 4 Best performing analyses when combining estimates across databases using meta-analysis, as measured by AUC, per outcome of interest

Acute liver failure	Acute myocardial infarction
AUC = .73 (SCC:409013) - Include first exposure only - Include first outcome only - Risk period: all time post exposure - Don't include index date in risk period - Control period: all time prior - Don't include index date in control period	AUC = .79 (SCCS:1907010) - Include all exposures - Prior: normal - Variance of prior: cross-validation - Risk period: all time post exposure - Don't include index date - Multivariate (include all drugs) - Require 0 days of observation time
Acute renal failure	**Upper GI bleeding**
AUC = .94 (SCC:409013) - Include first exposure only - Include first outcome only - Risk period: all time post exposure - Don't include index date in risk period - Control period: all time prior - Don't include index date in control period	AUC = .84 (SCCS:1901010) - Include all exposures - Prior: normal - Variance of prior: cross-validation - Risk period: 30 days from exposure start - Don't include index date - Multivariate (include all drugs) - Require 0 days of observation time

AUC area under the receiver operator characteristics curve, *SCC* Self Controlled Cohort, *SCCS* self-controlled case series

settings documented in Table 6 achieve fairly high performance both in EU-ADR and OMOP data. However, in Fig. 2 it is clear that the on-average best performing setting SCC: 409013 (dark red line) does not always generate the highest AUC within a particular database for a particular HOI. Notably, in Aarhus a better performance was achieved by employing a case-control design for acute liver injury, acute myocardial infarction and upper GI bleeding. On average, the AUC was 0.12 points higher for the highest performing method-settings combination within a database and HOI when compared to SCC: 409013. Although this might seem to indicate that performance can be improved by tailoring the analysis to the database and HOI, it should be noted that a large portion of the increase in AUC can be explained by random error caused by the low number of negative and positive controls and subsequent large variability in the AUC estimates. The reason that the same settings work fairly well across databases and HOIs could be that all HOIs are acute outcomes, and that all studied drug-HOI pairs are strongly powered. Furthermore, many settings such as the number of controls per case, the number of variables in the propensity score or the length of the control period do not seem to be very specific for the type of data or HOI.

In Table 6 we also see in general higher AUCs for the OMOP databases when compared to the EU-ADR databases. However, it should be noted that for every database we restricted the computation of the AUC to those test cases for which the database contained sufficient data (i.e. with statistical power to detect a relative risk of 1.25), and that therefore directly comparing the performance in the larger OMOP databases to the EU-ADR databases is like comparing apples to oranges: many more test-cases were used to compute the AUCs in the OMOP databases compared to the EU-ADR databases, and the overlap in test-cases is subsequently limited. We therefore focus on comparing the relative performance of methods within either OMOP or EU-ADR.

The results presented here, and those of the OMOP experiment, can be of help when interpreting the result of general drug safety studies. None of the methods and settings was found to consistently produce accurate estimates. As can be seen in Fig. 3, often relatives risks estimates of 2 or higher are observed when in fact no effect is present. One recommendation is to always include a set of negative control drugs (where no causal relationship is believed to

Table 5 Best performing analyses per method, measured as average AUC across all databases and outcomes. These analyses correspond to the lines shown in Fig. 2, and were used to produce the estimates shown in Fig. 3

Case–control	**Cohort method**	**Disproportionality**
AUC = 0.61 (CC:2000153)	AUC = 0.59 (CM:21001102)	AUC = 0.60 (DP:102009)
- Up to 10 controls per case	- Comparator: most prevalent drug with same indication, not in same class	- Include only first outcome
- Require 30 days observation time	- Washout period: 180 days	- Don't stratify by age and sex
- Risk period: all time post exposure	- Covariate eligibility: 180 days prior exposure	- Risk window: length of exp. +60 days
- Don't include index date	- Risk window: length of exposure +30 days	- Use BCPNN
- Match on age, sex, and visit (30 days)	- Covariate selection: HDPS	
- No nesting	- 100 top confounders from among	
- Include all exposures	- 200 most prevalent covariates	
	- That occur in at least 100 persons	
	- Trim upper and lower 5 %	
IC Temporal Pattern Discovery	**LGPS + LEOPARD**	**Self-controlled cohort design**
AUC = 0.67 (ICTPD:3054001)	AUC = 0.59 (LGPS: 18001007)	AUC = .75 (SCC:409013)
- Control period: 180 days prior exposure	- First exposure only	- Include first exposure only
- Risk period: 360 days from exposure start	- 365 day run-in period	- Include first outcome only
- Don't use 1 day before exposure	- No carry-over period	- Risk period: all time post exposure
- Don't use 1 m prior in expected	- Apply shrinkage	- Exclude index date from analysis
calc.	- LEOPARD filtering	- Control period: all time prior
Self-controlled case series		
AUC = .67 (SCCS:1963010)		
- Include all exposures		
- Prior: normal		
- Variance of prior: cross-validation		
- Risk period: all time post exposure		
- Include index date in risk window		
- Univariate (include one drug at a time)		
- Require 180 days of observation time		

AUC area under the receiver operator characteristics curve, *CC* case control, *DP* disproportionality methods, *ICTPD* Information Component Temporal Pattern Discovery, *LEOPARD* Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs, *LGPS* Longitudinal Gamma Poisson Shrinker, *SCC* Self Controlled Cohort, *SCCS* self-controlled case series

exist) in a study; If the magnitude of the control associations is non-trivial this can provide an insight on the residual bias that the data in the database cannot deal with, and this can be of help in interpreting the study findings. A potential further refinement could be to restrict the negative controls to those with the same indication as the drug of interest where we might expect the confounding to be more similar. For methods like CM we must consider that these have low AUCs, and the resulting estimates are prone to large error. If the methods presented here were to be used for automated signal generation, the computed ROC curves can be used to select a signaling threshold with the desired operating characteristics.

Although these recommendations are based on data of the six databases described here and the five databases in the OMOP experiment, it is unclear whether the findings are applicable to another database or other HOIs than the four studied here, or to drugs whose reaction takes place in a different timeframe with respect to the drugs in the 'positive control' set. Another limitation of this study includes the limited size of the databases, leading to low numbers of drug–outcome pairs with enough statistical power. This smaller sample size was also a consequence of the much lower incidence rates of the HOIs in EU-ADR databases compared to those in OMOP, which is most likely caused by the much broader HOI definitions used in OMOP. An investigation of the use of more narrow definitions in OMOP has shown that these do not lead to better method performance as measured in AUC [18], and we also do not see higher AUCs in EU-ADR with the narrower, arguably more specific HOI definitions than those in OMOP. Not all settings in the original experiment could be tested because the only diagnoses available were the HOIs. Furthermore, these studies are hampered by the lack of a true gold standard. Even though the reference set used here was created

Table 6 Analyses with the highest average AUC (averaged over outcomes and databases with enough data). Optimal settings according to the five databases in the most recent OMOP experiment and according to the six databases in this experiment are shown. The analyses for CM and CC in OMOP data could not be executed in EU-ADR data because these settings required conditions other than the health outcomes of interest to be included in the data. The optimal analyses for CM in EU-ADR data was a new analysis that was not yet included in the US OMOP experiment

	OMOP average overall optimal settings			EU-ADR average overall optimal settings		
Method	Settings ID	AUC in OMOP data	AUC in EU-ADR data	Settings ID	AUC in OMOP data	AUC in EU-ADR data
🟥 SCC	403002	0.81	0.72	409013	0.80	0.75
🟧 ICTPD	3054001	0.75	0.67	3054001	0.75	0.67
🟪 SCCS	1939010	0.71	0.64	1963010	0.69	0.67
🟩 CM	21000216	0.69		21001102		0.59
🟪 LGPS	18001024	0.58	0.56	18001007	0.54	0.59
🟥 CC	2000031	0.54		2000153	0.50	0.61
🟦 DP	101009	0.53	0.57	102009	0.52	0.60

AUC area under the receiver operator characteristics curve, *CC* case control, *DP* disproportionality methods, *ICTPD* Information Component Temporal Pattern Discovery, *LEOPARD* Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs, *LGPS* Longitudinal Gamma Poisson Shrinker, *SCC* Self Controlled Cohort, *SCCS* self-controlled case series

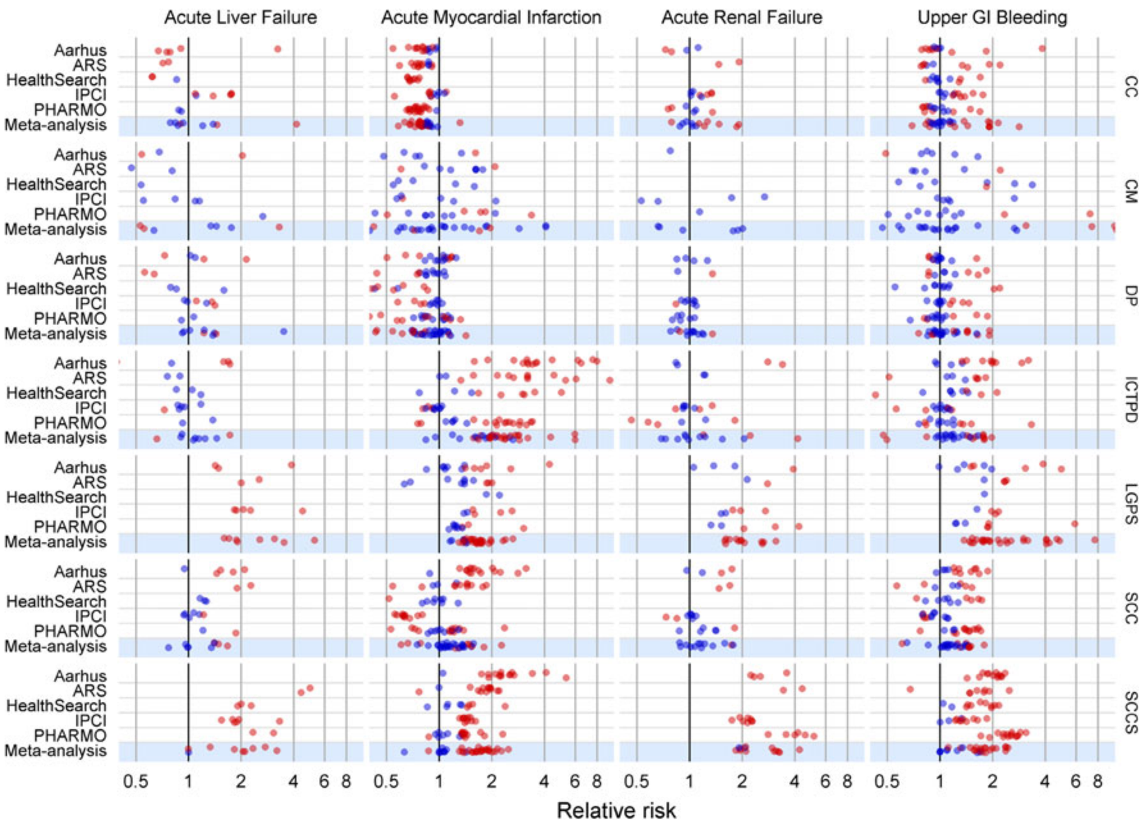

Fig. 3 Estimates for the negative control drugs, where the assumed true relative risk is one. For each method, those analysis choices were used that achieved on the highest area under the receiver operator characteristics curve (AUC) averaged over all databases and outcomes (see Table 4 for analysis details). *Red* indicates relative risks that are statistically significant different from 1 (two-sided $p < 0.05$)

meticulously by experts in the field, these experts had to rely on the limited data available, and it is possible that not all drug-HOI pairs were classified correctly as positive or negative control. 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 drugs, which could bias the results.

6 Conclusions

In conclusion, the result presented in this paper show great consistency with those observed in the recent OMOP experiment [3], giving us confidence that the recommendations we make are generalizable to similar databases.

Acknowledgments The authors wish to acknowledge the Observational Medical Outcomes Partnership (OMOP) investigators for their help in running the OMOP methods in the EU-ADR environment. The study was funded by The Foundation for the National Institutes of Health, grant STUR11OMOP. Dr. Schuemie received a fellowship from the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration, has become an employee of Janssen Research & Development since completing this research, and is an OMOP investigator. OMOP is funded by the Foundation for the National Institutes of Health 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. Mr. Straatman and Dr. Herings are employees of the PHARMO Institute for Drug Outcomes Research, which performs financially supported studies for several pharmaceutical companies. All other authors declare to have no conflicts of interests to declare. Dr. Sturkenboom is an employee of the Erasmus University Medical Center, and coordinates studies that are financially supported by several pharmaceutical companies (Novartis, Pfizer, EliLilly, Boehringer), none related to this study.

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. 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. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A comparison empirical performance of methods for a risk identification system. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0108-9.
4. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform. 2012;45(4):689–96.
5. Avillach P, Coloma PM, Gini R, Schuemie M, Mougin F, Dufour JC, et al. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. J Am Med Inform Assoc (JAMIA). 2013;20(1):184–92.
6. Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. N Engl J Med. 1988;318(6):352–5.
7. Madigan D, Schuemie MJ, Ryan PB. Empirical performance of the case-control method: lessons for developing a risk identification and analysis system. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0105-z.
8. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Zorych D. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. Drug Saf. 2013 (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 applid to health care databases. Drug Saf. 2013 (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. 2013 (in this supplement issue). doi:10.1007/s40264-013-0095-x.
11. Schuemie MJ, Madigan D, Ryan PB. Empirical performance of LGPS and LEOPARD: lessons for developing a risk identification and analysis system. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0107-x.
12. Ryan PB, Schuemie MJ, Madigan D. Empirical performance of a self-controlled cohort method: lessons for developing a risk identification and analysis system. Drug Saf. 2013 (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. 2013 (in this supplement issue). doi:10.1007/s40264-013-0100-4.
14. Hartzema AG, Reich CG, Ryan PB, Stang PE, Madigan D, Welebob E, et al. Managing data quality for a drug safety surveillance system. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0098-7.
15. 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.
16. Cantor SB, Kattan MW. Determining the area under the ROC curve for a binary diagnostic test. Med Decis Making. 2000;20(4):468–70.
17. 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.
18. Reich CG, Ryan PB, Schuemie MJ. Alternative outcome definitions and their effect on the performance of methods for observational outcome studies. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0111-1.