Drug Saf (2013) 36 (Suppl 1):S133–S142
DOI 10.1007/s40264-013-0107-x

ORIGINAL RESEARCH ARTICLE

Empirical Performance of LGPS and LEOPARD: Lessons for Developing a Risk Identification and Analysis System

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

Abstract

Background The availability of large-scale observational healthcare data allows for the active monitoring of safety of drugs, but research is needed to determine which statistical methods are best suited for this task. Recently, the Longitudinal Gamma Poisson Shrinker (LGPS) and Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs (LEOPARD) methods were developed specifically for this task. LGPS applies Bayesian shrinkage to an estimated incidence rate ratio, and LEOPARD aims to detect and discard associations due to protopathic bias. The operating characteristics of these methods still need to be determined.

Objective Establish the operating characteristics of LGPS and LEOPARD for large scale observational analysis in drug safety.

Research Design We empirically evaluated LGPS and LEOPARD in five real observational healthcare databases and six simulated datasets. We retrospectively studied the predictive accuracy of the methods when applied to a collection of 165 positive control and 234 negative control drug-outcome pairs across four outcomes: acute liver injury, acute myocardial infarction, acute kidney injury, and upper gastrointestinal bleeding.

Results In contrast to earlier findings, we found that LGPS and LEOPARD provide weak discrimination between positive and negative controls, although the use of LEOPARD does lead to higher performance in this respect. Furthermore, the methods produce biased estimates and confidence intervals that have poor coverage properties.

Conclusions For the four outcomes we examined, LGPS and LEOPARD may not be the designs of choice for risk identification.

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.

M. J. Schuemie (✉)
Department of Medical Informatics, Erasmus University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands
e-mail: m.schuemie@erasmusmc.nl

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

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

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

1 Background

Even though every drug has been extensively tested before being released on the market, there is a strong need for post-marketing drug safety monitoring. Traditionally, this has been done by relying on spontaneous reports of suspected adverse drug reactions [1, 2], but recent high-impact drug safety issues such as cardiovascular risk associated with rofecoxib has forced rethinking of the way we continuously assess the safety of drugs on the market [3]. One promising direction is to use observational health care data such as

electronic health records or insurance claims data for safety monitoring. In the past, such databases have been mostly used to confirm or refute potential safety issues originating from spontaneous reporting, but appropriate use of this data could also be invaluable in identifying new potential risks. To make this a reality, US Congress passed the Food and Drug Administration (FDA) Amendment Act, which called for the establishment of an "active post market risk identification and analysis system" with access to patient-level observational data from 100 million lives by 2012 [4]. It is envisioned that such a system would "use sophisticated statistical methods to actively search for patterns in prescription, outpatient, and inpatient data systems that might suggest the occurrence of an adverse event, or safety signal, related to drug therapy." [5] In Europe, the EU initiated the PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium, http://imi-protect.eu) and EU-ADR (Exploring and Understanding Adverse Drug Reactions. http://euadr-project.org) projects. Initial results of the EU-ADR project indicate that it is indeed feasible to use observational healthcare data to identify potential adverse reactions [6].

Several methods have been proposed for use in a risk identification system, but little empirical research exists to inform the expected operating characteristics of these approaches. One such method is the Longitudinal Gamma Poisson Shrinker (LGPS) [7], an adaptation of the Gamma Poisson Shrinker [8] to use on longitudinal observational data. LGPS compares the incidence rate during exposure to the background rate for all subjects, while adjusting for age and sex. An empirical Bayesian approach is used to shrink the resulting incidence rate ratio (IRR) estimate, based on the amount of available data. LGPS was proposed to be used in combination with another method, Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs (LEOPARD). By comparing the rate of prescription after the occurrence of an outcome of interest to the rate of prescription before, LEOPARD aims to detect protopathic bias, and is used as a filter to remove spurious signals that are generated by LGPS. The combination of LGPS and LEOPARD achieved the highest score in the OMOP Cup [7], where the objective was to identify which drugs caused which outcomes in a simulated dataset as measured by mean average precision. They were also amongst the best performing methods in a comparison of methods using European electronic health record data, with Area Under the receiver operator characteristics curve (AUC) as metric of the method's ability to distinguish positive from negative control drug-outcome pairs in a manually constructed reference set [6].

LGPS and LEOPARD were tested in five real observational healthcare databases and six simulated datasets, retrospectively studying the predictive accuracy of the method when applied to a collection of 165 positive controls and 234 negative controls across four outcomes: acute liver injury, acute myocardial infarction, acute kidney injury, and upper gastrointestinal bleeding. We estimated how well the methods can be expected to identify true effects and distinguished them from situations where there is no effect. We also explored the statistical properties of the LGPS effect estimates. With this empirical basis in place, LGPS and LEOPARD can be evaluated to determine whether they represent a potential tool to be considered in establishing a risk identification and analysis system to study the effects of medical products.

2 Methods

2.1 Overview of LGPS and LEOPARD

In the LGPS, we formulate the IRR as the ratio between the number of events observed during exposure O_1 and the expected number of events E:

$$IRR = \frac{O_1}{E}$$

E is computed by multiplying the duration of exposure with the incidence ratio when not exposed. In order to correct for age and sex, the expected count is first computed per age and sex strata, and then summed:

$$E = \sum_s t_{1s} \frac{O_{0s}}{t_{0s}}$$

where t_{1s} is the time exposed in strata s, O_{0s} is the number of events observed in strata s while not exposed, and t_{0s} is the unexposed time in strata s. In the current analysis, only the first occurrence of an outcome was considered, recurring outcomes were ignored. The reason for this is that in observational health care data it is often difficult to distinguish between a recurrent or 'new' outcome and follow-up of previous outcomes. Also, prior outcomes are usually strong risk factors for recurring outcomes of the same type, which could lead to several biases in the analysis.

To apply the shrinkage, each observed count O_1 is assumed to be drawn from a Poisson distribution with unknown mean μ and we are interested in the ratio $\lambda = \mu/E$. An empirical Bayesian model is fitted to all observed frequencies, and this model is used to determine the posterior distribution of each λ. We used the geometric mean of the posterior distribution as the estimated IRR.

LEOPARD is aimed at detecting protopathic bias, which transpires when a drug is prescribed to treat the disease itself or an early manifestation of a disease before the event is captured in the database. LEOPARD is based on the comparison of rates of drug prescriptions in the 25 days prior and 25 days after the occurrence of a condition. In case of a

true adverse drug reaction, we would expect the number of prescriptions prior to the event to be larger than the number of prescriptions after the event because the events are caused in part by the prescribed drug. In case of protopathic bias, because we are only looking at signals with increased relative risk there should be an increased rate of prescription prior to the event compared to baseline risk. However, this increase should continue or even be higher after the event since the drug in question is actually used to treat the disease and not cause it. We therefore assume that an increase in the number of prescriptions after an event relative to the number of prescriptions prior to the event is an indication of protopathic bias. A one-sided binomial test is used to test whether the number of prescriptions is greater in the post-event window. Signals where this test produces a p-value smaller than a predefined threshold are flagged as protopathic bias and are discarded. We used a threshold of 0.5, which was found to produce optimal results in previous simulation studies [7]. These studies also showed little sensitivity to the exact choice of threshold, since most p-values are either very small or very large.

Several analysis choices are required within LGPS and LEOPARD to enable a fully specific analysis. In this experiment, we varied five choices. They are illustrated in Fig. 1, and are described below:

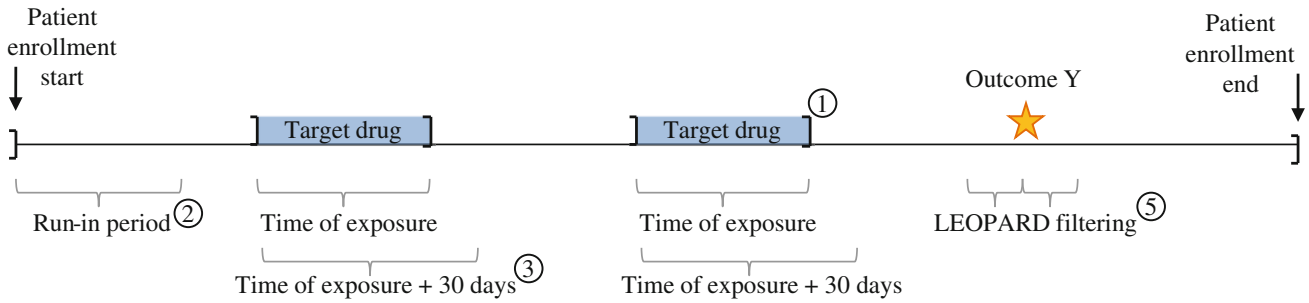
1. Should the analysis consider only the first period of persistent exposure, or should all exposures (including prevalence use) be included?
2. Run-in period: Should the first 365 days of patient data be included in the analysis, or should it be used only to capture patient history to be used for identifying incident users and outcomes?
3. Carry-over period: Should exposure be considered to be only the time on the drug, or should we add 30 days following the exposure to capture potential delayed effects of the drug?
4. Shrinkage: Should the Bayesian shrinkage be applied, or should the IRR (corrected only for age and sex) be used as the output?
5. LEOPARD: Should LEOPARD filtering be applied?

In this study, all 32 unique combinations of the five analysis choices were evaluated.

2.2 Experiment Design

The study was conducted against five 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 GE 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 [9] to model the MSLR database, and replicated six times to allow for injection of signals of known size (relative risk = 1, 1.25, 1.5, 2, 4, 10). The data used is described in more detail elsewhere [10].

The method was executed using all 32 analysis choice combinations against 399 drug-outcome pairs to generate an effect estimate and standard effort for each pair and parameter combination. These test cases include 165 'positive controls'—active ingredients with evidence to suspect a positive association with the outcome—and 234 'negative controls'—active ingredients with no evidence to



LGPS design parameters:

1. All exposures or first exposure only?
2. Run-in period: exclude first 365 days?
3. Carry-over period: add 30 days to risk window?
4. Use shrinkage (not shown)
5. Use pre and post outcome periods to estimate protopathic bias? (LEOPARD)

Fig. 1 Analysis choices within LGPS (Longitudinal Gamma Poisson Shrinker) and LEOPARD (Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs)

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 [11]. For every database we restricted the analysis to those drug-outcome pairs with sufficient power to detect a relative risk of RR <= 1.25, based on the age-by-gender-stratified drug and outcome prevalence estimates [12].

2.3 Metrics

The estimates and associated standard errors for all of the analyses are available for download at: http://omop.org/Research. To gain insight into the ability of a method to distinguish between positive and negative controls the IRR estimates were used to compute the AUC, a measure of predictive accuracy [13]: an AUC of 1 indicates a perfect prediction of which test cases are positive, and which are not. An AUC of 0.5 is equivalent to random guessing.

Often we are not only interested in whether there is an effect or not, but would also like to know the magnitude of the effect. However, in order to evaluate whether a method produces correct relative risk estimates, we must know the true effect size. In real data, this true effect size is never known with great accuracy for positive controls, and we must restrict our analysis to the negative controls where we assume that the true relative risk is 1. Fortunately, in the simulated data sets we do know the true relative risk for all injected signals. Using both the negative controls in real data, and injected signals in the simulated data, we compute the coverage probability: the percentage of confidence intervals that contain the true relative risk. In case of an unbiased estimator with accurate confidence interval estimation we would expect the coverage probability to be 95 %.

Lastly, we investigated to what degree each parameter can influence the estimated relative risk. For every parameter, we evaluated how much the estimated relative risk changed as a consequence of changing a single parameter while keeping all other parameters constant.

3 Results

3.1 Predictive Accuracy of all Settings

Figure 2 highlights the predictive accuracy, as measured by AUC, of all LGPS and LEOPARD design parameters across the four outcomes and five databases. Overall, LGPS and LEOPARD has poor performance, with AUCs around or below 0.5, the equivalent of random guessing. In general, performance improves when using LEOPARD filtering, although not in GE. With LEOPARD filtering, higher AUCs (>0.7) are achieved for acute myocardial infarction and upper GI bleeding.

For each outcome-database scenario we identified the analysis choices that yielded the highest AUC, as listed in Table 1. An optimal analysis (LGPS: 18001032) had the highest predictive accuracy for discriminating test cases for acute liver injury in CCAE (AUC = 0.58) and MDCR (AUC = 0.56), acute renal failure in CCAE (AUC = 0.61), acute myocardial infarction in CCAE (AUC = 0.71), GE (AUC = 0.58) and for upper GI bleeding in CCAE (AUC = 0.77). A different analysis (LGPS: 18001016) yielded the highest AUC in 5 outcome-database scenarios: acute liver injury in MDCD (AUC = 0.58), acute renal failure in MSLR (AUC = 0.61), acute myocardial infarction in MSLR (AUC = 0.61), GI bleed in MDCR (AUC = 0.77) and MSLR (AUC = 0.57). Most optimal analyses did not involve shrinkage, except for the GE database, where for 3 of the 4 outcomes the optimal analysis included shrinkage. LEOPARD filtering was always part of the optimal analysis, except for acute liver failure and acute renal failure in GE. The dashed and dotted lines in Fig. 2 indicate the performance of these top-performing analysis across databases for the same outcome, showing that the optimal analysis for one database can sometimes perform poorly when used in another database for the same outcome.

3.2 Overall Optimal Settings

The analysis with the best average performance across the 20 outcome-database scenarios is highlighted in the shaded grey line, and represents analysis LGPS:18001032. LGPS:18001032 is the unique identifier that reflects the analysis which uses first occurrences of drug exposure, a 365 day run-in period, 30 day carry-over period, no shrinkage, and applies LEOPARD filtering. This analysis was observed to perform well for all outcomes in CCAE, and was almost consistently among the least poor performing analyses for other outcomes. In the remainder of this paper we will use LGPS:18001032 as the representative analysis for LGPS in combination with LEOPARD.

"Appendix" contains the effect estimates for all test cases across the five databases using this optimal analysis (LGPS:18001032). To illustrate patterns in these findings, we discuss four specific test cases for upper GI bleeding, as shown in Fig. 3. Indomethacin is a non-steroidal anti-inflammatory drug (NSAID), and is known to be associated with upper GI bleeding [14]. An increased risk was therefore correctly identified in all five of the databases. Nabumetone is another nonsteroidal anti-inflammatory drug (NSAID) that is assumed to have a slightly lower risk of GI bleed [15]. However, the observed relative risk is much lower than reported in the literature, which might be due to gastro protective strategies such as co-prescribing with proton-pump

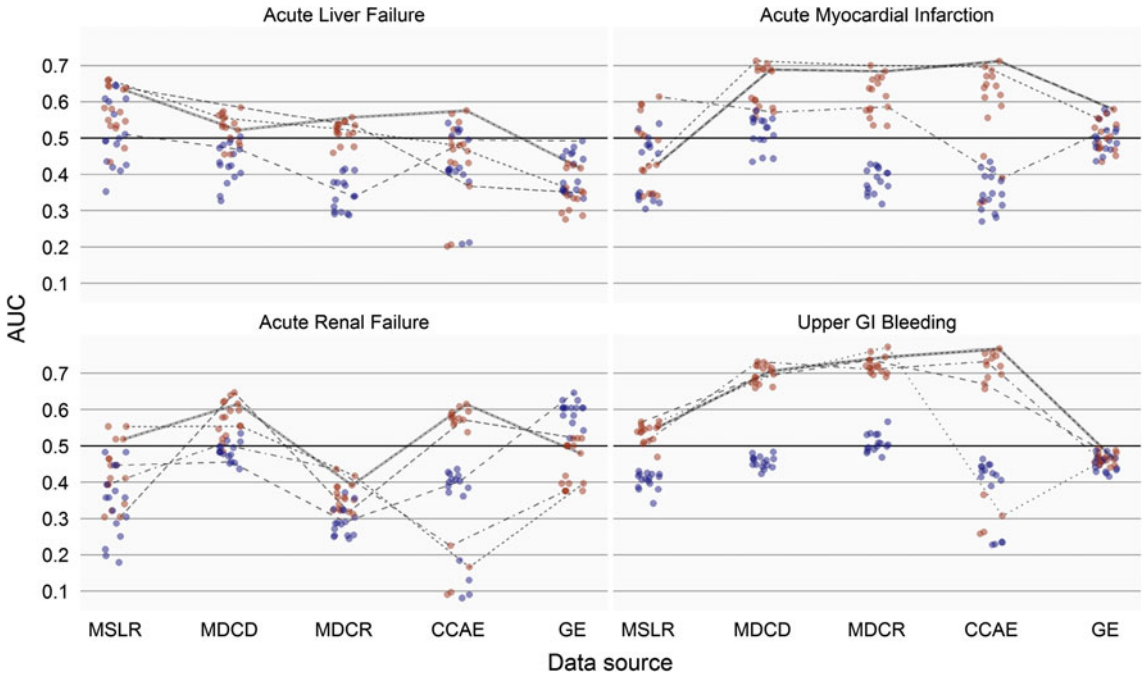


Fig. 2 Area under ROC Curve (AUC) for LGPS (Longitudinal Gamma Poisson Shrinker) and LEOPARD (Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs), by outcome and database. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity. *Each dot* represents one of the 32 unique analysis choice combination for LGPS and LEOPARD. *Red dots* indicate performance when LEOPARD filtering is used. The *solid grey line* highlights the analyses that had the highest average AUC across all 20 outcome-database scenarios. The *dashed lines* identify those analyses having the highest AUC in at least one database for a particular outcome

inhibitors. Prochlorperazine is a dopamine receptor antagonist that is often prescribed to alleviate nausea and vomiting caused by chemotherapy [16], and the high risk estimate for GI bleeding is therefore likely due to its association with cancer and with medication that is known to cause severe gastrointestinal distress. This bias was correctly detected by LEOPARD in four of the five databases. Metaxalone is a muscle relaxant that is not known to cause GI bleed, and is often prescribed as part of pain management in arthritis patients [17]. Its observed association could be due to its co-prescribing with drugs that do cause GI bleeds such as NSAIDs, or due to the fact that recipients of this drug are in general less healthy than the overall population. All associations found in the GE database were flagged as potential protopathic bias by LEOPARD. The main reason for this might be misclassification of the date of outcome or prescribing in this database.

3.3 Bias

Figure 4 shows the magnitude of bias observed across the estimates for the negative control test cases in the five real databases. The blue distributions show estimates for drug-outcome pairs not filtered by LEOPARD, the red distributions show estimates that are flagged as protopathic bias. We see across all four outcomes and all 5 databases that LGPS is positively biased, that is the expected value for the design when applied to a negative control is greater than 1. We also see that those estimates that were flagged as protopathic bias in general were indeed more positively biased. In GE, all negative control drugs for acute myocardial infarction were flagged as protopathic bias.

3.4 Coverage Probability

Figure 5 shows the coverage probabilities on simulated data. In general, the coverage is low, with the true effect size often falling below the estimated confidence interval. As the true effect size increased more often the true effect size was above the confidence interval, but not in any scenarios did the method achieve a coverage probability >28 %. For acute liver injury, when the true effect size is RR = 1 (that is, no signals injected), the coverage probability = 19 %, with the majority of the remaining 81 % of positive controls distributed below the estimated intervals. When we injected signals for the acute liver injury positive controls at RR = 2, the coverage probability increased to 24. When the true effect size was increased to RR = 10, the coverage probability was measured at 3 %.

Table 1 Optimal analysis choices for LGPS (Longitudinal Gamma Poisson Shrinker) and LEOPARD (Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs), by outcome and database

Data source	Acute liver injury	Acute renal failure	Acute myocardial infarction	Upper GI bleeding
CCAE	AUC = 0.58 (LGPS: 18001032)	AUC = 0.61 (LGPS: 18001032)	AUC = 0.71 (LGPS: 18001032)	AUC = 0.77 (LGPS: 18001032)
	1: First exposure only	1: First exposure only	1: First exposure only	1: First exposure only
	2: 365 day run-in period	2: 365 day run-in period	2: 365 day run-in period	2: 365 day run-in period
	3: 30 day carry-over period	3: 30 day carry-over period	3: 30 day carry-over period	3: 30 day carry-over period
	4: No shrinkage	4: No shrinkage	4: No shrinkage	4: No shrinkage
	5: LEOPARD filtering	5: LEOPARD filtering	5: LEOPARD filtering	5: LEOPARD filtering
GE	AUC = 0.49 (LGPS: 18001021)	AUC = 0.65 (LGPS: 18001029)	AUC = 0.58 (LGPS: 18001032)	AUC = 0.49 (LGPS: 18001015)
	1: First exposure only	1: First exposure only	1: First exposure only	1: First exposure only
	2: No run-in period	2: No run-in period	2: 365 day run-in period	2: 365 day run-in period
	3: No carry-over period	3: 30 day carry-over period	3: 30 day carry-over period	3: 30 day carry-over period
	4: Shrinkage	4: Shrinkage	4: No shrinkage	4: Shrinkage
	5: No LEOPARD filtering	5: No LEOPARD filtering	5: LEOPARD filtering	5: LEOPARD filtering
MDCD	AUC = 0.58 (LGPS: 18001016)	AUC = 0.65 (LGPS: 18001020)	AUC = 0.71 (LGPS: 18001024)	AUC = 0.73 (LGPS: 18001027)
	1: First exposure only	1: All exposure	1: First exposure only	1: All exposure
	2: 365 day run-in period	2: No run-in period	2: No run-in period	2: No run-in period
	3: 30 day carry-over period	3: No carry-over period	3: No carry-over period	3: 30 day carry-over period
	4: No shrinkage	4: No shrinkage	4: No shrinkage	4: Shrinkage
	5: LEOPARD filtering	5: LEOPARD filtering	5: LEOPARD filtering	5: LEOPARD filtering
MDCR	AUC = 0.56 (LGPS: 18001032)	AUC = 0.44 (LGPS: 18001008)	AUC = 0.70 (LGPS: 18001024)	AUC = 0.77 (LGPS: 18001016)
	1: First exposure only	1: First exposure only	1: First exposure only	1: First exposure only
	2: 365 day run-in period	2: 365 day run-in period	2: No run-in period	2: 365 day run-in period
	3: 30 day carry-over period	3: No carry-over period	3: No carry-over period	3: 30 day carry-over period
	4: No shrinkage	4: No shrinkage	4: No shrinkage	4: No shrinkage
	5: LEOPARD filtering	5: LEOPARD filtering	5: LEOPARD filtering	5: LEOPARD filtering
MSLR	AUC = 0.66 (LGPS: 18001007)	AUC = 0.55 (LGPS: 18001016)	AUC = 0.61 (LGPS: 18001016)	AUC = 0.57 (LGPS: 18001016)
	1: First exposure only	1: First exposure only	1: First exposure only	1: First exposure only
	2: 365 day run-in period	2: 365 day run-in period	2: 365 day run-in period	2: 365 day run-in period
	3: No carry-over period	3: 30 day carry-over period	3: 30 day carry-over period	3: 30 day carry-over period
	4: Shrinkage	4: No shrinkage	4: No shrinkage	4: No shrinkage
	5: LEOPARD filtering	5: LEOPARD filtering	5: LEOPARD filtering	5: LEOPARD filtering

AUC Area under ROC curve; Database abbreviations: *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity; Analysis choices: 1. Exposures to include, 2. Include run-in period prior to including subject, 3. Time-at-risk, 4. Apply Bayesian shrinkage, 5. Apply LEOPARD filtering

3.5 Parameter Sensitivity

Table 2 shows how sensitive effect estimates were to the specific analysis choices. Date source (the database used) has the largest effect on the estimates. The median change in effect estimates when changing from one database to another is 23 %. In other words, when holding all other analysis choices constant, there is a 50 % chance that the IRR observed in one database will change at least 28 % either positively or negatively when switching to another database. There is a 10 % chance that the impact of changing the database on the relative risk will be 80 % or more. The estimates were less sensitive to other analysis choices.

4 Discussion

In this paper we evaluated the LGPS and LEOPARD methods in the context of risk identification. LGPS is an adaption of the original GPS algorithm to longitudinal data, and

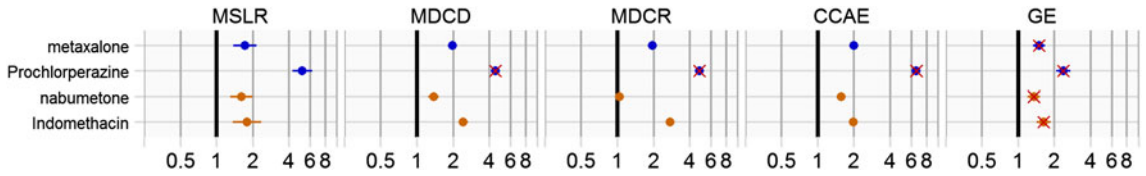


Fig. 3 IRR (incidence rate ratio) and 95 % confidence interval for 4 example drugs and upper gastrointestinal bleeding, across databases, using the overall optimal settings. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity; *Blue* negative controls; *Orange* positive controls; *Red crossed* indicate flagged as protopathic bias by LEOPARD. *Each line* represents point estimate and 95 % confidence interval for the drug-outcome pair in a particular database

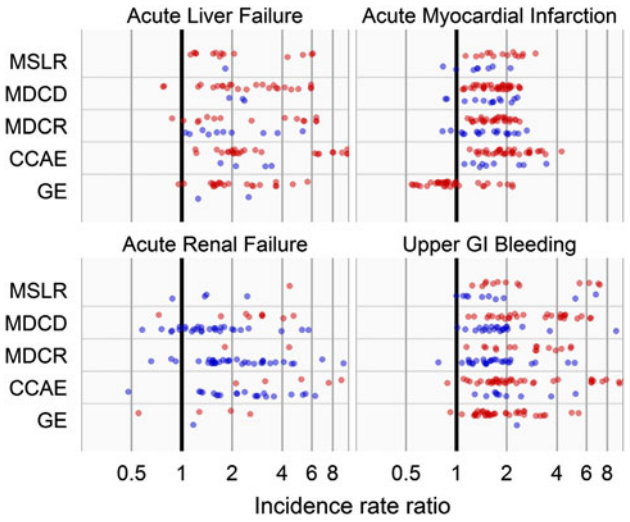


Fig. 4 Bias distribution: *Dots* show the estimates for negative controls, where the true incidence rate ratio is assumed to be 1. *Blue* distributions show drugs not filtered by LEOPARD (Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs), *red* distributions show drugs filtered by LEOPARD. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

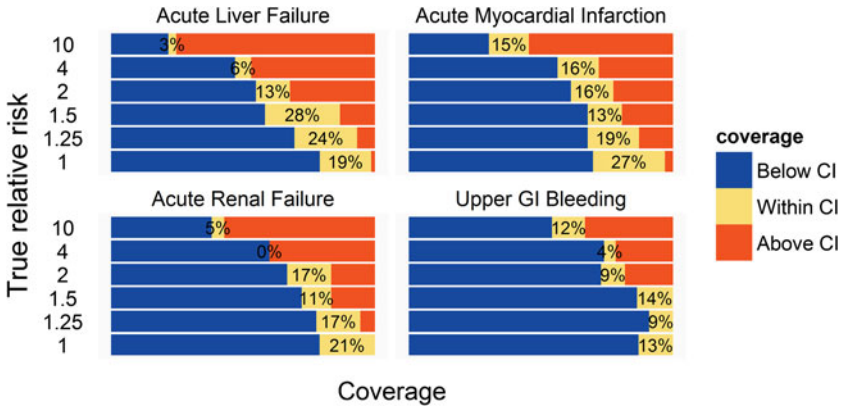
currently uses a population background rate as denominator when computing the IRR. As such, it is a simple method that could be prone to bias because it compares people that take drugs and are therefore more likely to have at least one ailment to what is in general a healthy population. Only confounding by age and sex is taken into consideration, and the LEOPARD method was proposed to eliminate at least one other type of bias, but certainly other types could remain. Despite these shortcomings, the combination of LGPS and LEOPARD has proven to perform well in discriminating positive from negative controls in the past, achieving the highest mean average precision in comparison to a large array of other methods on simulated data in the OMOP Cup [7], and in a recent evaluation using a network of European EHR databases LGPS and LEOPARD were amongst the best performing methods, achieving an AUC of 0.83 [6]. It is therefore surprising to see the low performance in this

Table 2 Sensitivity to analysis choices within GPS (Longitudinal Gamma Poisson Shrinker) and LEOPARD (Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs)

Parameter	q10 delta	q50 delta	q90 delta
Data source	1.04	1.23	1.80
First or all exposures	1.01	1.08	1.23
Run-in period	1.01	1.06	1.29
Carry-over period	1.01	1.05	1.26
Apply shrinkage	1.00	1.01	1.07

Q10/50/90 delta–10/50/90th percentile on the absolute change in point estimate observed across all outcome/database scenarios by holding all other parameters constant and changing the target parameter to an alternative value

Fig. 5 Coverage probability of LGPS (Longitudinal Gamma Poisson Shrinker) at different levels of true effect size, by outcome



evaluation. The best performance (AUC = 0.77) was achieved for upper GI bleeding using either CCAE or MDCR, but in several situations and for several outcomes, such as acute liver failure and acute renal failure, the performance is almost equivalent to random guessing.

There can be several explanations for this discrepancy with earlier findings. In the past, these methods have only been tested on either simulated data or European EHR data, while in this evaluation either insurance claims were used or the American EHR database GE. This latter database differs significantly from the European General Practitioner (GP) databases since in Europe every person is registered with one GP, and that GP is the gatekeeper of the patient's healthcare and therefore has a fairly complete overview of the health care received by the patient including in-patient care. The GE database on the other hand captures only partial information on the patient, and is likely to miss in-patient data. Another potential source of differences is the reference set used to evaluate the methods. Whereas the OMOP reference [11] set contains almost 100 controls per outcome, but for only four different outcomes, the EU-ADR reference set [18] contains only up to 10 controls per outcome, but for 10 different outcomes. The four outcomes included in OMOP are also included in the 10 in EU-ADR, and these seem to be more 'difficult' outcomes such as acute myocardial infarction and acute liver injury, where there is large potential for confounding by indication and other types of confounding. Some outcomes exclusive to EU-ADR, such as rhabdomyolysis and anaphylactic shock can be expected to be less prone to such bias. Also, for the four outcomes in common, a first glance does suggest that the OMOP reference set contains controls that can be expected to be more problematic in terms of bias. For instance, in the OMOP reference set, the negative controls for acute myocardial infarction included two antidiabetic drugs, while it is know that diabetes is a risk factor for myocardial infarction and therefore a confounder in this study. If it is indeed true that the OMOP reference set represents more difficult, confounding-riddled examples, then it should be more suitable to evaluate to what extent methods are able to deal with confounding and bias, whereas the EU-ADR reference set is more representative of situations where confounding is less likely and other aspects of a method, such as dealing with smaller sample sizes, are considered more important.

One finding that was consistent with previous studies is that in general the LEOPARD filter increased performance. The only exception was when LEOPARD was applied to the GE database, which could be explained by the incomplete data capture in this database. As noted earlier [6] LEOPARD seems informative about potential bias, but the variability in its performance should be taken into account. Using the outcome of LEOPARD as a binary filter, as was done here, might not be the best approach. We see that earlier evaluations of two-stage processes with a significance cutoff also showed poor performance [19]. Instead, it might be better to incorporate the output in a probabilistic framework. Bayesian shrinkage was typically not part of the optimal settings. The reason for this is probably the large size of the databases used. Since shrinkage only occurs when little data is available, it will not have an effect in large databases.

Previous evaluations of these methods only focused on the ability to distinguish between positive and negative controls. In this study, we also investigated whether the magnitude of the effect found by the method was correct, and found that LGPS was positively biased, overestimating the effect size of both negative controls in real data and positive controls in simulated data. This finding is hardly surprising, since the method compares people on drugs to a generally healthy background population.

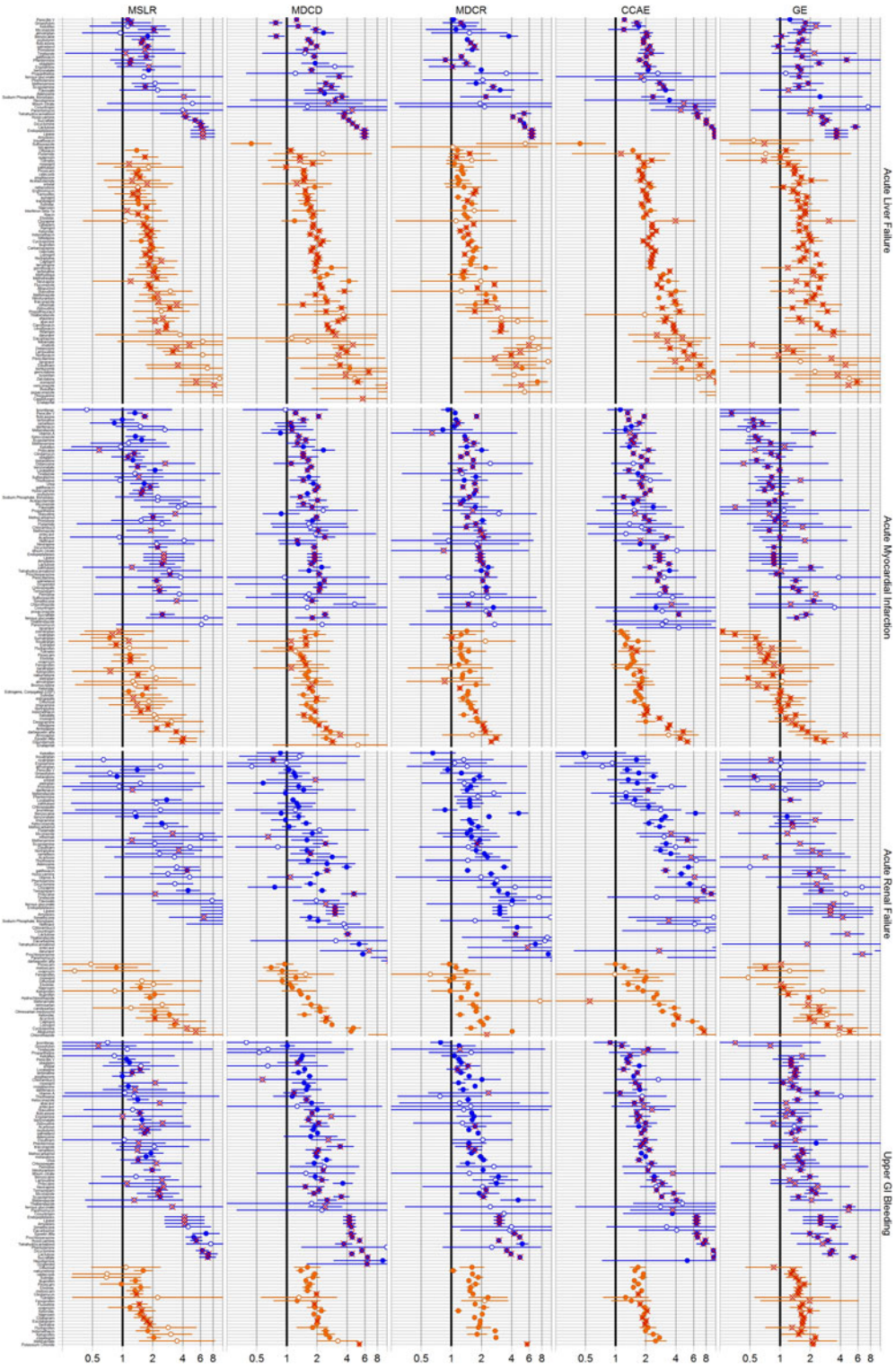
5 Conclusions

The results presented here cast doubts on the generalizability of earlier findings that suggested that LGPS and LEOPARD are well suited for risk identification in longitudinal observational data. Certainly, more research is needed to understand the differences in performance found in this study and previous studies.

Acknowledgements The Observational Medical Outcomes Partnership 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. Drs. Schuemie and Ryan 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. Dr. Madigan has 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 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.

Appendix



LGPS and LEOPARD 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; *Red crosses* indicate flagged by LEOPARD as protopathic bias. *Open circles* indicate a drug-outcome pair with not enough power to detect a minimal detectable rate of 1.25. *Each line* represents point estimate and 95 % confidence interval for the drug-outcome pair in a particular database

References

1. Ahmad SR. Adverse drug event monitoring at the Food and Drug Administration. J Gen Intern Med. 2003;18(1):57–60.
2. Olsson S. The role of the WHO programme on International Drug Monitoring in coordinating worldwide drug safety efforts. Drug Saf. 1998;19(1):1–10.
3. Avorn J. Evaluating drug effects in the post-Vioxx world: there must be a better way. Circulation. 2006;113(18):2173–6.
4. Public Law 110-85: Food and Drug Administration Amendments Act of 2007. 2007.
5. Woodcock J, Behrman RE, Dal Pan GJ. Role of postmarketing surveillance in contemporary medicine. Ann Rev Med. 2011;62:1–10.
6. 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.
7. Schuemie MJ. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. Pharmacoepidemiol Drug Saf. 2011;20(3):292–9.
8. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat. 1999;53(3):190–6.
9. Ryan PB, Schuemie MJ. Evaluating performance of risk identification methods through a large-scale simulation of observational data. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0110-2.
10. Overhage JM, Ryan PB, Schuemie MJ, Stang PE. Desideratum for evidence based epidemiology. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0102-2.
11. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0097-8.
12. 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.
13. Cantor SB, Kattan MW. Determining the area under the ROC curve for a binary diagnostic test. Med Decis Mak. 2000;20(4):468–70.
14. Masso Gonzalez EL, Patrignani P, Tacconelli S, Garcia Rodriguez LA. Variability among nonsteroidal antiinflammatory drugs in risk of upper gastrointestinal bleeding. Arthritis Rheum. 2010;62(6):1592–601.
15. Ashworth NL, Peloso PM, Muhajarine N, Stang M. Risk of hospitalization with peptic ulcer disease or gastrointestinal hemorrhage associated with nabumetone, Arthrotec, diclofenac, and naproxen in a population based cohort study. J Rheumatol. 2005;32(11):2212–7.
16. Gralla RJ, Osoba D, Kris MG, Kirkbride P, Hesketh PJ, Chinnery LW, et al. Recommendations for the use of antiemetics: evidence-based, clinical practice guidelines. American Society of Clinical Oncology. J Clin Oncol. 1999;17(9):2971–94.
17. Richards BL, Whittle SL, Buchbinder R. Muscle relaxants for pain management in rheumatoid arthritis. Cochrane Database Syst Rev. 2012;1:CD008922.
18. Coloma PM, Avillach P, Salvo F, Schuemie MJ, Ferrajolo C, Pariente A, et al. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. Drug Saf. 2013;36(1):13–23.
19. Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. Stat Med. 1989;8(12):1421–32.