Drug Saf (2013) 36 (Suppl 1):S123–S132
DOI 10.1007/s40264-013-0106-y

ORIGINAL RESEARCH ARTICLE

Evaluation of Disproportionality Safety Signaling Applied to Healthcare Databases

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

Abstract

Objective To evaluate the performance of a disproportionality design, commonly used for analysis of spontaneous reports data such as the FDA Adverse Event Reporting System database, as a potential analytical method for an adverse drug reaction risk identification system using healthcare data.

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

Measures 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 the design generates. The primary measure was the area under the curve (AUC) of the receiver operating characteristic (ROC) curve.

Results For each combination of 4 outcomes and 5 databases, 48 versions of disproportionality analysis (DPA) were carried out and the AUC computed. The majority of the AUC values were in the range of $0.35 < \mathrm{AUC} < 0.6$, which is considered to be poor predictive accuracy, since the value $\mathrm{AUC} = 0.5$ would be expected from mere random assignment. Several DPA versions achieved AUC of about 0.7 for the outcome Acute Renal Failure within the GE database. The overall highest DPA version across all 20 outcome-database combinations was the Bayesian Information Component method with no stratification by age and gender, using first occurrence of outcome and with assumed time-at-risk equal to duration of exposure + 30d, but none were uniformly optimal. The relative risk estimates for the negative control drug-event combinations were very often biased either upward or downward by a factor of 2 or more. Coverage probabilities of confidence intervals from all methods were far below nominal.

Conclusions The disproportionality methods that we evaluated did not discriminate true positives from true negatives using healthcare data as they seem to do using spontaneous report data.

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.

W. DuMouchel
Oracle Health Sciences, Burlington, Massachusetts, USA

W. DuMouchel (✉)
Oracle Health Sciences Global Business Unit,
7164 N Mercer Spring Pl, Tucson, AZ 85718, USA
e-mail: bill.dumouchel@oracle.com

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

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

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

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

1 Background

Disproportionality analysis (DPA) methods for drug safety surveillance represent the primary class of analytic methods for analyzing data from spontaneous report systems (SRSs). SRSs receive reports that comprise of one or more drugs, one or more adverse events (AEs), and possibly some basic demographic information (in addition to narrative and text data). A report typically contains fields such as date of report, age and gender of patient, as well as a list of drugs to which the patient has been recently exposed, and a list of adverse event names [typically drawn from the Medical Dictionary for Regulatory Activities (MedDRA) structured vocabulary] that the patient has experienced. The number of drugs and the number of events reported can range from 1 up to dozens, while the total number of drugs and event names in the search space is of course many thousands. Disproportionality analysis methods include the multi-item gamma-Poisson shrinker (MGPS), proportional reporting ratios (PRR), reporting odds ratios (ROR), and Bayesian confidence propagation neural network (BCPNN). The methods search SRS databases for "interesting" associations and focus on low-dimensional projections of the data, specifically 2-dimensional contingency tables. Table 1 shows a typical table.

DPA methods compute a measure of association for each such table, and also use some threshold value to declare a "signal" if the measure exceeds the threshold. MGPS focuses on the "reporting ratio" (RR). The RR for the drug *i*—adverse event *j* combination (RR_{ij}) is the observed number of occurrences of the combination (20 in the example below) divided by the expected number of occurrences. MGPS computes the expected value under a model of independence. Specifically, in the example above, overall, drug *i* occurs in 10 % of the reports and adverse event *j* occurs in 10 % of the reports. Thus, if drug *i* and adverse event *j* are statistically independent, 0.1*0.1 = 1 % of reports should include both drug *i* and AE *j*, that is 12 reports in this case. Thus the RR for this example is 20/12 or 1.67; this combination occurred about 67 % more often than expected.

The common measures of association between drug *i* and event *j* are of the form N_{ij}/E_{ij}, where $N_{ij} = a$ for the *(i, j)* version of Table 1 and E_{ij} is a comparator value or *Expected Count* intended to represent the null hypothesis of no association. Definitions of expected counts *E* and disproportionalities for four measures are shown in Table 2.

Table 1 A fictitious 2-dimensional projection of an SRS database

Report contains	Drug i Yes	Drug i No	Total
Event j	*a*	*b*	120
Yes	20	100	
Event j	*c*	*d*	1,080
No	100	980	
Total	120	1,080	1,200

In the Stratified Reporting Ratio, the value of *E* is computed separately for each set of reports within a set of exclusive and exhaustive strata—for example all combinations of 5 age groups, 2 genders, and 10 report year groupings, giving 100 strata indexed by *s*, each having its own E_s, which are summed, as recommended for a covariate-adjusted 2 × 2 table analysis by Mantel–Haenszel [1].

Drugs that cause a particular adverse event will typically receive a higher score than drugs that don't. Conversely, if an adverse event and a drug are stochastically independent, all measures will return the null hypothesis value of 1, subject to sampling variability. This statistical variability diminishes as the sample size increases. In the SRS context, however, the count in the $N = a$ cell is often small, leading to substantial variability (and hence uncertainty about the true value of the measure of association) despite the often large numbers of reports overall. PRR and ROR do not address the variability issue whereas MGPS and BCPNN adopt a Bayesian approach to reduce the effect of random noise. MGPS places a prior distribution on RRs that encapsulates a prior belief that most RRs are close to the average value of all RR's (i.e., close to 1). Only in the face of substantial evidence from the data does MGPS return an RR estimate that is substantially larger than one. Thus, for example, an RR of 100 that derives from an observed count of $N = 1$ might result in a MGPS RR estimate (Empirical Bayesian Geometric Mean or EBGM) of 2 (i.e. the crude RR is shrunk towards a value of 1) whereas an RR of 100 that derives from an observed count of $N = 100$ might result in a EBGM RR estimate of 95. The EBGM score is the geometric mean of the posterior distribution of the true RR. Other summaries are possible. For example, Almenoff et al. [2], mentions "EB05", which is equivalent to the quantity $\lambda_{0.05}$ in DuMouchel and Pregibon [3]. This is the 5th percentile of the posterior distribution—meaning that, according to the Bayesian analysis, there is a 95 % probability that the "true" RR exceeds the EB05. Since EB05 is always smaller than EBGM this, in a sense, adds extra shrinkage and represents a more conservative choice than EBGM.

As shown in the above display, the formula for the Bayesian Information Component is a very simple modification of the non-Bayesian version: just add 0.5 to both N and E. This corresponds to the "True RR" having a Gamma(1/2, 1/2) prior distribution. The prior distribution for MGPS is estimated from the database as a whole and involves more complicated calculations [3, 4, 5].

In this study, we evaluated the performance of a disproportionality design, commonly used for analysis of spontaneous reports data such as FDA Adverse Event Reporting System (FAERS), as a potential analytical method for a risk identification system using healthcare data. We tested the

Table 2 Expected value definitions used by various disproportionality statistics

Comparison or expected value E	Disproportionality statistics based on this definition of E

(a + b)(a + c)/(a + b + c + d)	RR = N/E
	IC = $\log_2$(RR)
	BIC = $\log_2$ [(N + 0.5)/(E + 0.5)]

$\Sigma_s(a_s + b_s)(a_s + c_s)/(a_s + b_s + c_s + d_s)$	SRR = N/E
b(a + c)/(b + d)	PRR = N/E
bc/d	ROR = N/E

RR reporting ratio, *IC* information component, *BIC* Bayesian information component, *SRR* stratified reporting ratio, *PRR* proportional reporting ratio, *ROR* reporting odds ratio

disproportionality design in five real observational healthcare databases and 6 simulated datasets, retrospectively studying the predictive accuracy of the method when applied to a collection of 165 positive controls and 234 negative controls across four 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 the design generates. With this empirical basis in place, the disproportionality design can be evaluated to determine whether it represents a potential alternative tool to be considered in establishing a risk identification and analysis system to study the effects of medical products.

2 Methods

2.1 Applying DPA to Longitudinal Data

In the context of spontaneous report systems, some authors use the term "signal of disproportionate reporting" (SDR) when discussing associations highlighted by DPA methods [6, 7]. In reality, most SDRs that emerge from spontaneous report databases represent non-causal effects because the reports are associated with treatment indications (i.e., confounding by indication), co-prescribing patterns, co-morbid illnesses, protopathic bias, channeling bias, or other reporting artifacts, or, the reported adverse events are already labeled or are medically trivial. In this sense, SDRs *generate* hypotheses. Furthermore, spontaneous report databases present a number of well documented limitations such as under-reporting, over-reporting, and duplicate reporting. They fail to provide a denominator—how many individuals are actually consuming the drug, and generally have limited temporal information with regard to duration of exposure and the time order of exposure and condition [4, 6, 8]. The richer context of longitudinal data (such as claims databases or electronic health records) affords the possibility of more refined analysis to address some of these artifacts. Nonetheless, given the wide acceptance of DPA methods in pharmacovigilance, application of DPA methods to longitudinal data may prove useful. A key step in the application of DPA methods to any data is the mapping of the data into drug-condition two-by-two tables. Our approach to this mapping is to try to mimic what SRS reports the longitudinal data would generate.

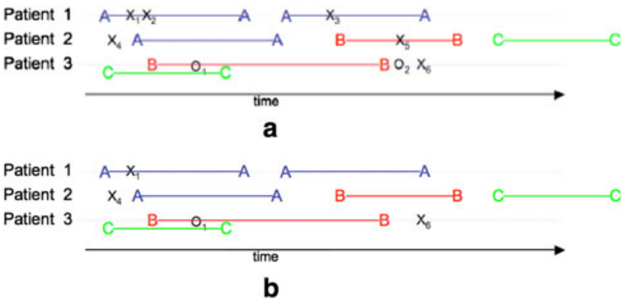


Fig. 1 **a** A longitudinal dataset with three patients, three distinct drugs (*A*, *B*, and *C*) and two distinct conditions (*X* and *O*). **b** A longitudinal dataset with three patients, three distinct drugs (*A*, *B*, and *C*) and two distinct conditions (*X* and *O*). Incident conditions only

We illustrate our approach using the example of Fig. 1a. Figure 1a shows three patients. Patient 1 consumed drug A during two separate drug eras. The patient experienced condition X three times during these eras, twice during the first era and once during the second. Patient 2 also had three drug eras but with three separate drugs, A, B, and C. Finally Patient 3 had two overlapping drug eras, one with drug B and one with drug C. The patient experienced condition O while taking both B and C, and conditions O and X after the drug eras. Note we treat conditions as if they occur at distinct moments in time. In fact the data may contain condition "eras" and what we are utilizing is the timestamp of the beginning of the era. Drug eras, on the other hand, play an important role in our approach. A drug era represents a continuous period of drug usage, possibly augmented with an additional off drug period. We refer to the optional off-drug period as a surveillance window and discuss this further below. In practice, defining the on-drug portion of the drug era itself requires design decisions. For example, should two 30-day prescriptions with a one-day gap between the two prescriptions be considered one drug era or two?

Consider how to construct the 2 × 2 table for drug A and condition X. N = *a* is the number of distinct X conditions that occur during drug A eras. *c* is the number of distinct non-X conditions that occur during drug A eras. *b* is the number of distinct X conditions that occur during non-A drug eras. d is the number of distinct non-X conditions that occur during non-A drug eras. Thus, for the example of Fig. 1a, *a* = 3 ($A + X_1$, $A + X_2$, $A + X_3$), *c* = 0, *b* = 1 ($B + X_5$), and *d* = 2 ($B + O_1$,$C + O_1$).

Our application of DPA to longitudinal data also makes a distinction between incident and prevalent conditions. The incident case only considers the first occurrence of each event, whereas the prevalent case (considered in the above example) considers all occurrences. Thus, for the example above, the incident analysis would proceed as above but only consider the first event of each type. Figure 1b illustrates the modified dataset used in an incident analysis. (Note that our use of the term "incident" does not necessarily coincide with standard use in epidemiological practice. In particular, we do not require an event-free "clean" period prior to first condition occurrence.) In Fig. 1b, the 2 × 2 table would be constructed as $a = 1\,(A + X_1), c = 0, b = 0$, and $d = 2\,(B + O_1, C + O_1)$.

There are other possible ways of constructing 2 × 2 tables, but preliminary results suggest that the method described above works at least as well as others we have tried [9]. Note that even within the SRS world more than one counting method has been used. If we consider Patient 3 in the above figures, a question arises as to whether the two drug-event pairs $(B + O_1)$, $(C + O_1)$ should count as $d = 2$ or $d = 1$, since they refer to a single patient experiencing a single event. According to the GPS method in DuMouchel [4] and PRR described in Evans et al. [10], $d = 2$, but using the methods of BCPNN [11] and MGPS [3], $d = 1$. As stated above, for this paper we will count all drug-event combinations and thus use $d = 2$.

2.2 Experimental Design

The study was conducted against five observational healthcare databases to allow evaluation of performance across different populations and data capture processes: MarketScan^{TM} Lab Supplemental (MSLR, 1.2 million (m) persons), MarketScan^{TM} Medicare Supplemental Beneficiaries (MDCR, 4.6 m persons), MarketScan^{TM} Multi-State Medicaid (MDCD, 10.8 m persons), Truven MarketScan^{TM} Commercial Claims and Encounters (CCAE, 46.5 m persons), and the GE Centricity^{TM} (GE, 11.2 m persons) database. GE is an EHR database; the other four databases contain administrative claims data. A 10 m-person simulated dataset was also constructed using the OSIM2 simulator [12] to model the MSLR database, and replicated 6 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 [12, 13]. The method was executed using 48 parameter combinations against 399 drug-outcome pairs to generate an effect estimate and standard error for each pair and parameter combination. The 48 parameter sets are all combinations of the following factors:

- Outcomes to include: First (incident) only, all occurrences
- Disproportionality metric: BCPNN (Bayesian Information Component), MGPS (Empirical Bayes Geometric Mean), Proportional Reporting Ratio (PRR)
- Stratification: no stratification, stratified by age and gender
- Time at risk: start of exposure up to 30 days, all observed time after start of exposure, exposed time plus up to 30 days after, exposed time plus up to 60 days after

The 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 [14]. For every database we restricted the analysis to those drug-outcome pairs with sufficient power to detect a relative risk of 1.25, based on the age-by-gender-stratified drug and outcome prevalence estimates. There are 234 negative controls and 165 positive controls. The estimates and associated standard errors for all of the analyses are available for download at: http://omop.org/Research.

2.3 Metrics

To gain insight into the ability of a method to distinguish between positive and negative controls the IRR estimates were used to compute the Area Under the receiver operator characteristics Curve (AUC), a measure of predictive accuracy [15]: 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%.

3 Results

3.1 Predictive Accuracy of All Settings

Figure 2 highlights the predictive accuracy, as measured by AUC, of the 48 disproportionality design parameters that were used for each of the 4 outcomes and 5 databases. For each outcome-database scenario we identified the parameter settings that yielded the highest AUC, as listed in Table 3. An optimal setting (labeled DP: 101009 in Table 3) had the highest predictive accuracy for discriminating test cases for acute kidney injury in MDCR. (AUC = 0.42) Note that none of the DP variations was able to even rise to the level of random guessing for the two outcomes Acute Liver Failure or Acute Renal Failure on the MDCR database. But the parameter settings that did the best for Acute Renal Failure on that database was also the one that had the best average AUC across all 20 outcome-database combinations. Its description is shown in italics in Table 3, namely (a) first occurrence of outcome only, (b) BCPNN/BIC metric, (c) no age-sex stratification, (d) time-at-risk equals length of exposure + 30d. The performance of these parameter settings on all 20 outcome-database applications is indicated by the solid lines on Fig. 2. They show that these settings are among the best performing on perhaps half of the 20 combinations, but are only in the middle range of accuracy on many of the other database-outcome combinations.

Perhaps the most striking aspect of Fig. 2 is how poorly all the disproportionality variations perform across these outcome-database combinations. For 5 of the 20 combinations, no DP method even approaches the random-guessing level of AUC = 0.5, and in only 3 of the combinations can even the highest-scoring DP method surpass AUC = 0.61. Only the Acute Renal Failure-GE data combination has a respectably high cluster of AUC scores, ranging from about 0.62 to 0.77.

The dashed and dotted lines in Fig. 2 indicate the performance of the best-scoring setting across outcomes within the same database, showing that the optimal setting for one outcome often performs poorly when used for another outcome in the same database.

3.2 Overall Optimal Settings

As mentioned above, the overall best scoring DP method used: (a) first occurrence of outcome only, (b) BCPNN/BIC metric, (c) no age-sex stratification, and d) time-at-risk equal length of exposure + 30d. In the remainder of this paper we will use these as the representative settings for the disproportionality method.

"Appendix" contains the effect estimates for all test cases across the 5 databases using this optimal parameter setting (DP: 101009). To illustrate patterns in these findings, we discuss four specific test cases for acute liver injury, as shown in Fig. 3.

One drug known to be causative agent is isoniazid, which was used as a positive control. The association between isoniazid and acute liver injury was consistently one of the largest effects observed using the disproportionality design, with all 5 databases generating IRR between 3 and 5 with very tight confidence intervals, as shown in Fig. 3. In contrast,

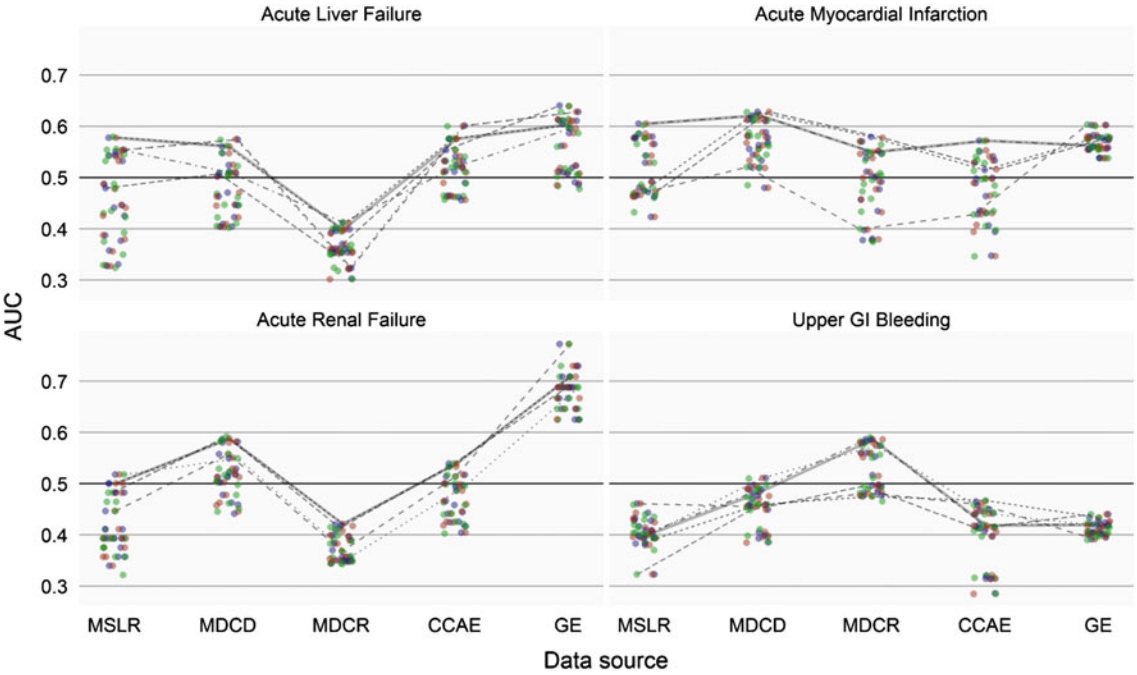


Fig. 2 Area under ROC Curve (AUC) for disproportionality parameters, by outcome and database. Each *dot* represents one of the 48 parameter combinations of the disproportionality method. *Dot colors*: *blue*: PRR, *red*: BCPNN, *green*: MGPS. The *solid grey line* highlights the parameter that had the 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. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE centricity

Table 3 Optimal disproportionality parameter settings for each outcome and database. Italic indicates the setting with the highest average AUC across all outcomes and databases

Source	Acute liver injury	Acute kidney injury	Acute myocardial infarction	Upper GI bleeding
CCAE	AUC = 0.60 DP: 103008) Outcomes to include: first occurrence Metric: PRR Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: all time post-exposure start	AUC = 0.54 (DP: 101008) Outcomes to include: first occurrence Metric: PRR Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: length of exposure + 30d	AUC = 0.57 (DP: 101008) Outcomes to include: first occurrence Metric: PRR Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: length of exposure + 30d	AUC = 0.47 (DP: 103003) Outcomes to include: all occurrences Metric: MGPS Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: all time post-exposure start
GE	AUC = 0.64 (DP: 103001) Outcomes to include: all occurrences Metric: PRR Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: all time post-exposure start	AUC = 0.77 (DP: 103001) Outcomes to include: all occurrences Metric: PRR Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: all time post-exposure start	AUC = 0.60 (DP: 107003) Outcomes to include: all occurrences Metric: MGPS Stratify by age: yes Stratify by gender: yes Stratify by year: no Time-at-risk: All time post-exposure start	AUC = 0.44 (DP: 103008) Outcomes to include: first occurrence Metric: PRR Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: All time post-exposure start
MDCD	AUC = 0.57 (DP: 103009) Outcomes to include: first occurrence Metric: BCPNN Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: all time post-exposure start	AUC = 0.59 (DP: 101010) Outcomes to include: first occurrence Metric: MGPS Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: length of exposure + 30d	AUC = 0.63 (DP: 105008) Outcomes to include: first occurrence Metric: PRR Stratify by age: yes Stratify by gender: yes Stratify by year: no Time-at-risk: length of exposure + 30d	AUC = 0.51 (DP: 101001) Outcomes to include: all occurrences Metric: PRR Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: length of exposure + 30d
MDCR	AUC = 0.41 (DP: 104003) Outcomes to include: all occurrences Metric: MGPS Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: 30d from exposure start	*AUC = 0.42 (DP: 101009)* *Outcomes to include: first occurrence* *Metric: BCPNN* *Stratify by age: no* *Stratify by gender: no* *Stratify by year: no* *Time-at-risk: length of exposure + 30d*	AUC = 0.58 (DP: 105008) Outcomes to include: first occurrence Metric: PRR Stratify by age: yes Stratify by gender: yes Stratify by year: no Time-at-risk: length of exposure + 30d	AUC = 0.59 (DP: 101010) Outcomes to include: first occurrence Metric: MGPS Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: length of exposure + 30d
MSLR	AUC = 0.58 (DP: 101010) Outcomes to include: first occurrence Metric: MGPS Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: length of exposure + 30d	AUC = 0.52 (DP: 104008) Outcomes to include: first occurrence Metric: PRR Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: 30d from exposure start	AUC = 0.61 (DP: 101008) Outcomes to include: first occurrence Metric: PRR Stratify by age: no Stratify by gender: no Stratify by year: no Time-at-risk: length of exposure + 30d	AUC = 0.46 (DP: 107001) Outcomes to include: all occurrences Metric: PRR Stratify by age: yes Stratify by gender: yes Stratify by year: no Time-at-risk: all time post-exposure start

AUC area under the receiver operator characteristics curve, *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE centricity

erythromycin is another drug thought to be associated with acute liver injury and used as a positive control, but the IRR estimates for erythromycin-acute liver injury were significantly less than IRR = 1 in all 5 databases. The disproportionalities for erythromycin seem to be biased significantly downward, as shown in Fig. 3. Just as we would anticipate that positive controls should yield large and statistically significant findings, we desire negative controls to

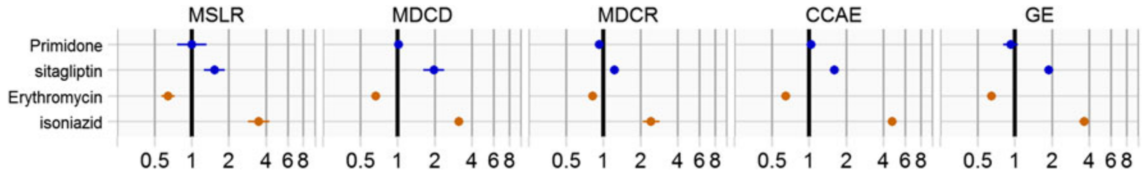


Fig. 3 Relative reporting rate and 95 % confidence interval for 4 example drugs and acute liver injury, across databases, using the overall optimal disproportionality settings. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE centricity. *Blue* negative controls, *orange* positive controls, each *line* represents point estimate and 95 % confidence interval for the drug-outcome pair in a particular database

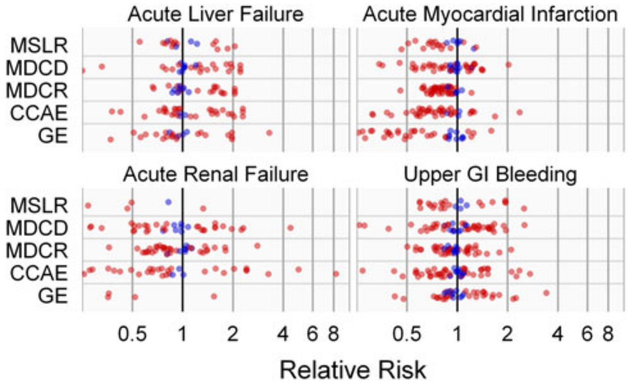


Fig. 4 Bias estimates for the negative control drugs, where the assumed true relative risk is one, using those settings that achieved the highest AUC averaged over all databases and outcomes. *Red* indicates relative risks that are statistically significant different from 1. *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE centricity

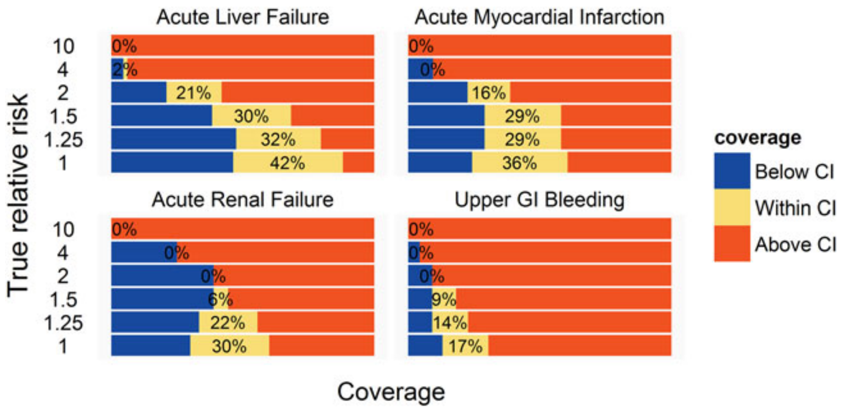
produce non-significant findings near the null value of IRR = 1. Sitagliptin is an anti-diabetic medication classified as a negative control due to lack of evidence of any association with acute liver injury. Unfortunately, Fig. 3 shows that the sitagliptin estimates are consistently between 1 and 2, with very tight confidence intervals that exclude the no-effect value of 1. The disproportionality estimates seem to have a positive bias for acute liver injury in all 5 databases. Another negative control, primidone, is an anticonvulsant that has not previously been associated with acute liver injury. Figure 3 shows that in this case all 5 disproportionality estimates are dead on the presumably correct value of 1.

So for these 4 drugs, the estimation of association with acute liver failure seems to be correct for 2 and incorrect for the other 2. Discouragingly, for the two drugs with incorrect estimates (sitagliptin and erythromycin) the disproportionality method had very tight confidence intervals—it was very sure of the wrong answer—which only confirms the low values of AUC shown in Fig. 2.

3.3 Bias

Figure 4 shows the magnitude of bias observed for the disproportionality method across the estimates for the negative control test cases in the five real databases. There are 20 scatter plots in Fig. 4, showing the distribution of point estimates in each of the 20 HOI-database combinations. From Fig. 4, estimates for Acute Liver Failure and Upper GI Bleeding seem to be unbiased across all negative control drugs, since the distributions are fairly well centered on top of the no-effect value of 1. (Although the point estimates seem well centered about 1, many of the individual drug CIs still fail to cover 1). The spread of point estimates is not so well centered for Acute Renal Failure and Myocardial Infarction. Probably the DP point estimates for negative control drugs are negatively biased for these two HOIs. The scatter plots in Fig. 4 show that the point estimates for negative controls are all over the map, belying their claimed tight confidence intervals. Results from

Fig.5 Coverage probability of disproportionality analysis (IC) at different levels of true effect size, by outcome



simulations (see Fig. 5) indicate that, when the null hypothesis is true, excessive variance is more of a problem than bias, and contrary to Fig. 4, it is Upper GI Bleed that shows the greatest negative bias.

3.4 Coverage Probability

Figure 5 shows the coverage probabilities for 95 % confidence intervals based on the highest-average-AUC disproportionality method on simulated data. As the figure shows, the degree of coverage falls woefully below 95 % for all four outcomes and across the full range of true relative risk. The degree of coverage decreased as the true effect size increased, with an increasing proportion of true effects falling above the upper bound. In no scenarios did the method achieve a coverage probability >42 %. Except for the Upper GI Bleeding outcome, the errors are roughly symmetric (equally positive and negative) for low true relative risk, but the estimates rapidly become underestimates for all outcomes as the true relative risk increases. At true RR = 10, 100 % of estimated upper 95 % limits fall below the true value.

4 Discussion and Conclusion

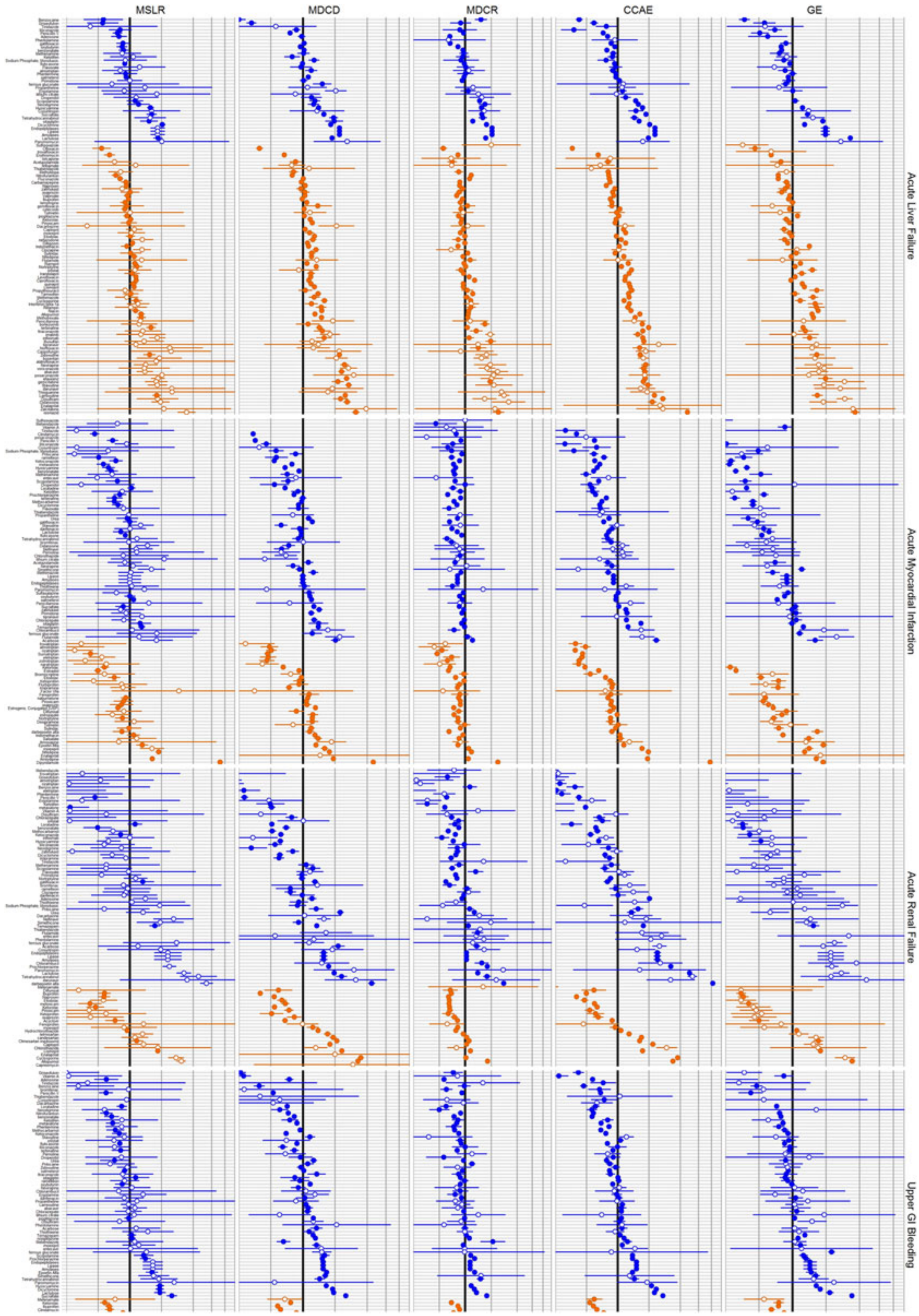
The results presented above seem to show that adapting disproportionality methods, designed for analysis of spontaneous report databases, and applying them to longitudinal healthcare data may be unfruitful. Our implementation of such a strategy has been singularly unimpressive, at least for the purpose of detecting true associations from among our nearly 400 drug-event OMOP “gold standard” associations. Figure 2 is a telling summary of how poorly the disproportionality methodology works, at least as implemented here. The four health outcomes of interest crossed with the 5 databases provide 20 separate test beds. In only one, acute renal failure within the GE data, was there a reasonably powerful detection capability (AUC about 0.7). Perhaps six or so other HOI-Database combinations approached AUC about 0.6, depending on which variation of disproportionality methodology was used. But the majority of such test beds could hardly exceed chance in their discriminatory power. Based on the disproportionality methods, Figs. 4 and 5 show that relative risk estimates and confidence limits are highly variable and unreliable.

How to explain these results? When applied to spontaneous reports databases such as FAERS, these and similar methods work well [16–21], routinely achieving AUC values from 0.7 up to 0.85 or so [22]. The EU-ADR project [23] also applies several DPA methods to a different set of databases using a different set of “gold standard” drug-event combinations. Their values of AUC averaged about 0.7 for the 4 versions of DPA that they report, whereas they report larger values, from 0.73 to 0.79, for other methods based on cohort analyses, case-control, or self-controlled case studies. This our results agree with theirs that DPA methods seem to be inferior when longitudinal health records are available. The values of AUC in our Fig. 2 are often even less than the by-chance value of 0.5, a surprising result that the EU-ADR did not observe, and for which we have no ready explanation. One possible explanation for why DPA methods work well for spontaneous report databases is that a spontaneous report of a suspected drug-ADR association has the benefit of a focused interpretation by the reporter, whereas the data in a healthcare database, being collected for a different purpose such as insurance billing, may tend to be subject to many more random sources of noise and bias. Since spontaneous report data lacks measures of total exposure to serve as a denominator for drug-ADR counts, the disproportionality methods designed for such data cannot take advantage of the fact that longitudinal health data does have such exposure measures (perhaps not perfectly ascertained). So we should not be surprised if other statistical estimation methods designed for healthcare databases provide improved power. Note that this critique of DPA methods is not a criticism of the particular mathematical algorithms applied to 2 × 2 tables like Table 1. Rather, the problem is that those tables only use numerator data (event counts) without being adjusted for exposure measures. At any rate, we do not recommend these disproportionality methods for analysis of drug-ADR associations in longitudinal healthcare data.

Acknowledgments The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health (FNIH) through generous contributions from the following: Abbott, Amgen Inc., AstraZeneca, Bayer Healthcare Pharmaceuticals, Inc., Biogen Idec, Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Janssen Research and Development, Lundbeck, Inc., Merck & Co., Inc., Novartis Pharmaceuticals Corporation, Pfizer Inc, Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi-aventis, Schering-Plough Corporation, and Takeda. Dr. DuMouchel is an employee of Oracle Health Sciences. 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 received grants from FNIH.

This article was published in a supplement sponsored by the Foundation for the National Institutesof Health (FNIH). The supplement was guest edited by Stephen J.W. Evans. It was peer reviewed byOlaf H. Klungel who received a small honorarium to cover out-of-pocket expenses. S.J.W.E hasreceived travel funding from the FNIH to travel to the OMOP symposium and received a fee fromFNIH for the review of a protocol for OMOP. O.H.K has received funding for the IMI-PROTECTproject.from the Innovative Medicines Initiative Joint Undertaking (http://www.imi.europa.eu) under GrantAgreement no 115004, resources of which are composed of financial contribution from theEuropean Union’s Seventh Framework Programme (FP7/2007-2013) and EFPIA companies’ in kindcontribution.

Appendix



Disproportionality 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 %

References

1. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959;22(4):719–48.
2. Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJ, Yuen N. Novel statistical tools for monitoring the safety of marketed drugs. Clin Pharmacol Ther. 2007;82(2):157–66.
3. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. San Francisco: ACM; 2001. p. 67–76.
4. Dumouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat. 1999;53(3):177–90.
5. Norén GN, Hopstadius J, Bate A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. Stat Methods Med Res 2013;22(1):57–69.
6. Hauben M, Madigan D, Gerrits CM, Walsh L, Van Puijenbroek EP. The role of data mining in pharmacovigilance. Expert Opin Drug Saf. 2005;4(5):929–48.
7. Hauben M, Reich L. Safety related drug-labelling changes: findings from two data mining algorithms. Drug Saf. 2004;27(10): 735–44.
8. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf. 2002;25(6):381–92.
9. Zorych I, Madigan D, Ryan P, Bate A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res. 2013;22(1):39–56.
10. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10(6): 483–6.
11. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54(4):315–21.
12. Ryan PB, Schuemie M. 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.
13. Overhage JM, Ryan PB, Schuemie MJ, Stang PE. Desideratum for evidence based epidemiology. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0102-2.
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. Cantor SB, Kattan MW. Determining the area under the ROC curve for a binary diagnostic test. Med Decis Making. 2000;20(4): 468–70.
16. Fram DM, Almenoff JS, DuMouchel W. Empirical Bayesian data mining for discovering patterns in post-marketing drug safety. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC: ACM; 2003. p. 359–68.
17. Almenoff JS, DuMouchel W, Kindman LA, Yang X, Fram D. Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. Pharmacoepidemiol Drug Saf. 2003;12(6):517–21.
18. DuMouchel W, Smith ET, Beasley R, Nelson H, Yang X, Fram D, et al. Association of asthma therapy and Churg-Strauss syndrome: an analysis of postmarketing surveillance data. Clin Ther. 2004;26(7):1092–104.
19. Almenoff JS, LaCroix KK, Yuen NA, Fram D, DuMouchel W. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. Drug Saf. 2006;29(10):875–87.
20. Solomon R, Dumouchel W. Contrast media and nephropathy: findings from systematic analysis and Food and Drug Administration reports of adverse effects. Invest Radiol. 2006;41(8): 651–60.
21. DuMouchel W, Fram D, Yang X, Mahmoud RA, Grogg AL, Engelhart L, et al. Antipsychotics, glycemic disorders, and life-threatening diabetic events: a Bayesian data-mining analysis of the FDA adverse event reporting system (1968-2004). Ann Clin Psychiatry. 2008;20(1):21–31.
22. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clin Pharmacol Ther. 2013;93(6):539–46.
23. Schuemie MJ, Coloma PM, Straatman H, et al. Using electronic healthcare records for drug safety signal detection; a comparative evaluation of statistical methods. Med Care. 2012;50(10):890–7.