Drug Saf (2013) 36 (Suppl 1):S33–S47
DOI 10.1007/s40264-013-0097-8

ORIGINAL RESEARCH ARTICLE

Defining a Reference Set to Support Methodological Research in Drug Safety

Patrick B. Ryan · Martijn J. Schuemie · Emily Welebob · Jon Duke · Sarah Valentine · Abraham G. Hartzema

Abstract

Background Methodological research to evaluate the performance of methods requires a benchmark to serve as a referent comparison. In drug safety, the performance of analyses of spontaneous adverse event reporting databases and observational healthcare data, such as administrative claims and electronic health records, has been limited by the lack of such standards.

Objectives To establish a reference set of test cases that contain both positive and negative controls, which can serve the basis for methodological research in evaluating methods performance in identifying drug safety issues.

Research Design Systematic literature review and natural language processing of structured product labeling was performed to identify evidence to support the classification of drugs as either positive controls or negative controls for four outcomes: acute liver injury, acute kidney injury, acute myocardial infarction, and upper gastrointestinal bleeding.

Results Three-hundred and ninety-nine test cases comprised of 165 positive controls and 234 negative controls were identified across the four outcomes. The majority of positive controls for acute kidney injury and upper gastrointestinal bleeding were supported by randomized clinical trial evidence, while the majority of positive controls for acute liver injury and acute myocardial infarction were only supported based on published case reports. Literature estimates for the positive controls shows substantial variability that limits the ability to establish a reference set with known effect sizes.

Conclusions A reference set of test cases can be established to facilitate methodological research in drug safety. Creating a sufficient sample of drug-outcome pairs with binary classification of having no effect (negative controls) or having an increased effect (positive controls) is possible

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/s40264-013-0097-8) contains supplementary material, which is available to authorized users.

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

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

J. Duke
Indiana University School of Medicine, Indianapolis, IN, USA

J. Duke
Regenstrief Institute, Indianapolis, IN, USA

S. Valentine · A. G. Hartzema
College of Pharmacy, University of Florida, Gainesville, Florida, USA

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

and can enable estimation of predictive accuracy through discrimination. Since the magnitude of the positive effects cannot be reliably obtained and the quality of evidence may vary across outcomes, assumptions are required to use the test cases in real data for purposes of measuring bias, mean squared error, or coverage probability.

1 Background

Drug safety assessment encompasses all scientific and data gathering activities relating to the detection, assessment, and understanding of adverse events of medical products through the product lifecycle [1]. Findings from randomized clinical trials, observational studies, and spontaneous adverse event reporting may all contribute to the evidence base for the safety profile of a product, but each source has its own unique limitations. A key challenge faced by the medical community is determining when the evidence is sufficiently compelling to influence the belief of whether or not the drug has a causal effect on the outcome. The 'truth' about a causal effect is generally unobtainable, but the question becomes "when can evidence about an association be appropriately used to draw a causal inference [2]?" One component necessary to addressing this question is a full understanding of the validity of the evidence that is generated. When we see evidence from a particular source, how often is that evidence indicative of a positive causal effect and how often could that evidence be observed when no such effect exists? In an ideal world, guidance that was grounded in empirical evidence would be available for evaluating information for all sources, such that observing specific levels of information would be known to provide definitive support of the causal effect and other findings would be known to be sufficient for supporting definitive conclusions that an effect could be ruled out.

In practice, when evaluating drug safety issues, findings from any one source are rarely definitive in either direction. Randomized, placebo-controlled clinical trials are often thought of as the gold standard in estimating true causal effects, and evidence from trials is highly regarded when available. However, most clinical trials apply specific inclusion/exclusion criteria for patient selection to improve the study's internal validity, but may limit the generalizability of study findings to the general population. Moreover, randomized trials are often underpowered for studying adverse events which are not frequently occurring, and often follow patients for a limited duration which may not reflect the entire time-at-risk [3]. Evidence from non-randomized observational studies and spontaneous adverse event reporting can complement knowledge gained from clinical trials, but the degree to which the evidence from these sources is consistent with 'truth' is not well understood.

Part of the challenge in evaluating the consistency of evidence with 'truth' is first establish a 'ground truth' set that can be used to retrospectively measure the performance of specific information sources. As data mining algorithms for spontaneous adverse event reporting databases were developed, several efforts to define a 'ground truth' reference set were performed to support methods evaluation. Lindquist et al. evaluated the performance of the Bayesian Confidence Propagation Neural Network based on the Martingdale and Physician Desk Reference compendium of drug information [4]. Hauben and Reich [5] compared the performance of proportional reporting ratio (PRR) and the multi-item gamma Poisson shrinker (MGPS) algorithms by defining positive controls based on label changes observed on MedWatch in 6 months in 2001. Hochberg et al. [6], selected 27 drugs and classified adverse events based on level of evidence from product labeling and literature review, and used this reference event database to evaluate three algorithms.

There is increasing interest in the use of observational healthcare databases, such as administrative claims and electronic health records, as part of an "active postmarket risk identification and analysis system". This interest has motivated the systematic application of existing designs [7–10] as well as the development of new statistical methods [11–13] and prompted the need for large-scale evaluation of the performance of observational analyses. The Exploring and Understanding Adverse Drug Reactions (EU-ADR) project conducted a methodologic experiment that examined the predictive accuracy of three classes of observational methods [14]. To enable their work, the team identified up to 5 positive and negative controls for 10 outcomes based on literature, spontaneous reporting, and clinical adjudication [15].

In its initial experiments, the Observational Medical Outcomes Partnership (OMOP) evaluated the performance of eight different methods using 53 drug-outcome pairs which were classified as 9 positive controls and 44 negative controls on the basis of product labeling and expert consensus [16, 17]. These test cases reflected a variety of relationships between the drug and outcome: outcomes with high or low background prevalence in the population (such as myocardial infarction vs. acute liver failure); outcomes that would likely occur acutely after drug administration (e.g., angioedema) versus those expected to have a delayed onset. Although the studies using the 53 drug-health outcome of interest [18] pairs provided valuable insights, a primary finding from this methodological work was that a much larger pool of test cases is necessary to more fully explore the interaction between data and analysis, and how performance may vary by drugs and health outcomes of interest. In this manuscript we describe the construction of a broader set of test cases for use in drug safety methodological research.

2 Methods

Four health outcomes of interest were selected for this analysis: acute myocardial infarction (AMI), acute kidney injury (AKI), acute liver injury (ALI), and upper gastrointestinal bleeding (UGIB). Four outcomes were chosen from among the original 10 outcomes studied in past OMOP experiments due to stakeholder interest in restricting focus on fewer outcomes to allow for deeper exploration within each outcome. The specific outcomes were selected because they have shown to be priority events of importance to pharmacovigilance activities [19], and each reflects different challenges to drug safety surveillance. Acute myocardial infarction is an outcome with a high background rate in the general population, and for which most observed events are not attributable to medical product exposure. Products associated with AMI commonly have events observed during randomized trials in development, but residual concerns remain due to the modest effect size (e.g. relative risk <2) which the trials are often not adequately powered to detect. Acute kidney injury is an important outcome for post-market drug surveillance since the kidney is a primary pathway for drug elimination, and while renal function is extensively monitored in pre-clinical and phase 1 studies, patients with pre-existing renal conditions are often excluded during late-stage drug development. AKI is inconsistently defined, and drug-induced cases can be difficult to distinguish among patients with renal impairment or progressing with chronic kidney disease, so pharmacovigilance activities commonly rely on both spontaneous reporting and strong pharmacological knowledge. Acute liver injury is regarded as a sentinel adverse event due to the liver's role in metabolizing drugs and the seriousness of consequences if hepatic function is impaired. Hepatotoxicity is the main reason for post-market product withdrawal, but because the incidence of drug-induced liver injury is low and proper diagnosis can be challenging [20], spontaneous adverse event reporting has been the primary mechanism to detect drug-related effects. Upper gastrointestinal bleeding is an outcome that has been extensively studied in pharmacoepidemiology, in part because drug-induced ulceration is sufficiently prevalent and the consequences are often serious enough to warrant healthcare encounters that are captured in many observational databases.

For each outcome, we attempted to define a broad set of positive controls, or drugs that are suspected to have a causal effect on the outcome, and a broad set of negative controls, or drugs for which there is no evidence of a causal relationship. This classification is based on available evidence from product labeling and systematic review of the literature. Figure 1 illustrates the criteria imposed to define the positive and negative control test cases.

To identify candidate positive controls, product labeling information was extracted from 5,333 structured product labels (SPLs) available from the FDA's DailyMed website through January 22nd, 2011 (http://dailymed.nlm.nih.gov). Duke et al. developed a software tool, known as the Structured Product Label Information Coder and Extractor (SPLICER) that uses natural language processing (NLP) to extract adverse event data from each section within the SPLs [21]. A previous study of its performance on 100 labels showed a recall of 92.8 % and a precision of 95.1 % [22]. All adverse events extracted from SPLs were mapped into Medical Dictionary for Regulatory Activities (MedDRA®) terms. For each outcome, we produced a list of potential MedDRA terms that would fall within the outcome definition [23]. Drugs were identified as candidate positive controls if one or more of the MedDRA terms for an outcome appeared in either a Black Box Warning or was

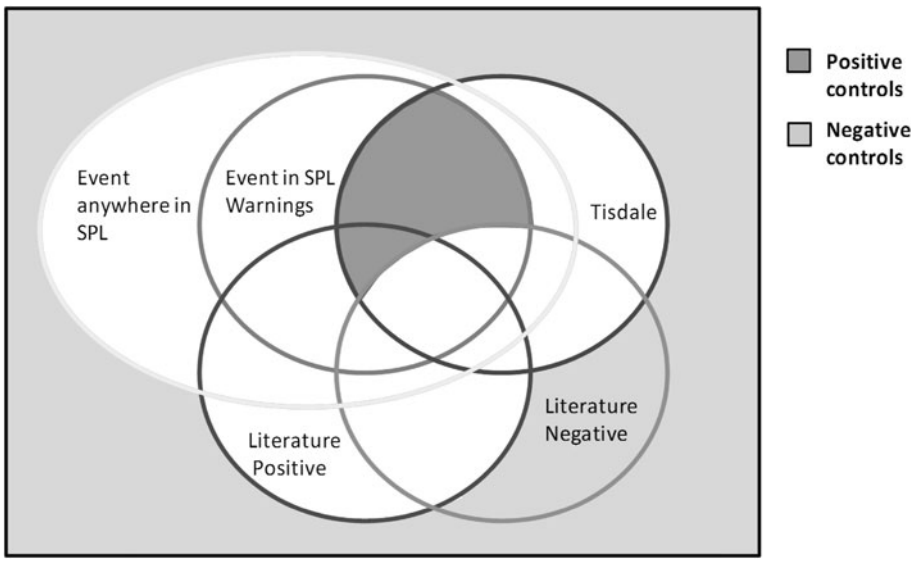

Fig. 1 Inclusion/exclusion criteria for positive and negative controls. SPL refers to the Structured Product Label of the drug of interest, Tisdale refers to Tisdale's literature review. Positive literature indicates the set of cases with at least one article confirming the existence of a causal relationship. Negative literature indicates the set of cases with at least one published study that was sufficiently powered but found no relationship between the drug and outcome

present in both the 'Warnings and Precautions' section and 'Adverse Reactions' section of a product label for the active ingredient. The candidate list was further refined through manual review to minimize misclassification from the automated processes. All candidate positive controls were evaluated at the active ingredient level, without consideration of differential effects at different dose levels or in specific formulations. Events attributed to drug–drug interactions were excluded. Additional details about the natural language processing for adverse event label extraction is provided elsewhere [24].

We compared the candidate positive controls identified through product labeling with those identified as causative agents from the systematic literature reviews provided by Tisdale et al. [25]. These reviews also provide a grading ('A', 'B', or 'C') of the level of evidence used to support the assessment of the product being considered a causative agent, where 'A' indicates evidence from one or more randomized controlled clinical trials, 'B' signifies support from nonrandomized clinical trials, prospective observational studies, cohort studies, retrospective studies, case-control studies, meta-analyses, and/or postmarketing surveillance studies, and 'C' cites evidence from one or more published case reports or case series. All levels of evidence were sufficient for preserving the candidate as a positive control test case. We used Tisdale's review of causative agents of angina pectoris, myocardial ischemia, and acute coronary syndrome to filter the list of candidate labeled test cases for acute myocardial infarction. To assess acute kidney injury, we used the assessment of drug-induced kidney injury resulting from renal hemodynamic alterations, acute tubular necrosis, acute interstitial nephritis, nephrolithiasis, or glomerulonephritis. Tisdale's review of causative agents for hepatic and cholestatic disease provided the basis for evaluating drugs that are associated with liver injury. The Tisdale review used for upper gastrointestinal bleeding included causative agents for upper gastrointestinal ulceration and esophageal damage, but excluded drugs with non-specific bleeding risks such as anticoagulants and antiplatelet medications.

We (SV, AH) conducted an independent literature review using PubMed® identified randomized trials and observational studies published from 1990 through 2011 whose primary outcome was one of the four adverse events of interest. This search was conducted to allow us an additional attempt to identify conflicting evidence that would potentially cast doubt on the classification of a drug-outcome relationship, and to allow us to extract study-specific information from each publication. Initial PubMed searches used the following Medical Subject Headings (MeSH) terms: 'Drug-Induced Liver Injury', 'Acute Kidney Injury', 'Myocardial infarction', 'Gastrointestinal Hemorrhage', each using MeSH subheadings of 'chemically induced', 'etiology', and 'epidemiology'. These were followed by narrow searches based on specific drugs or drug classes in combination with the outcomes in an attempt to identify related articles. The information gathered from each article included type of study, drug class and/or specific drug, confounding variables and control methods, latency period, and association measures (point estimate and associated 95 % confidence interval).

The independent literature review was not used to identify candidate positive controls, only to restrict the list of candidates that previously arose from product labeling and the Tisdale review. Drugs were excluded from consideration as positive controls if one or more studies produced conflicting evidence, as defined by a point estimate of the relative risk ≤ 1 (negative literature). Candidate positive controls based on product labeling and the Tisdale review for which we did not identify any published studies in our independent review were not excluded. True protective effects of products were not eligible as candidates for positive controls, because only increased risks were considered. The results section touches upon the consistency of effect estimates among positive controls with multiple published studies.

The same information sources (product labeling, Tisdale review, literature search) were used to define negative control test cases where there was no available evidence to suggest a causal effect of the drug on the outcome. We evaluated all active ingredients from the SPLs as candidate negative controls. We eliminated all products with an occurrence of any condition within the MedDRA high-level term of the target outcome in any section of any product label. For example, to be considered a negative control for 'acute myocardial infarction', a product could not contain any condition related to 'Ischaemic coronary artery disorders', such as 'acute coronary syndrome', 'myocardial ischemia', 'chest pain', and 'unstable angina'. We further excluded candidate negative controls if they were listed as causative agents in the Tisdale reviews for the outcomes. Finally, we eliminated drugs for which we identified conflicting evidence in the published literature, based on one or more randomized trials or population-based observational studies with point estimate >1 (positive literature). Case reports were not deemed sufficient evidence of positive literature to warrant exclusion of a negative control. Among those drug-outcome pairs meeting all negative control criteria, we sampled those drugs which were either classified as a positive control for a different outcome or satisfied the negative control criteria for at least three outcomes.

We restricted our positive and negative control test cases to those products that have at least one person exposed in each of the five observational databases that OMOP licensed for its experiments: MarketScan® Lab

Table 1 All test cases in the reference set; Drugs in **bold** have minimum detectable relative risk ≤1.25 in all 5 databases

Acute kidney injury				
Positive controls				
Acyclovir	**Naproxen** [37–39, 54]	Captopril	Etodolac	Moexipril
Hydrochlorothiazide	**Olmesartan medoxomil**	Chlorothiazide	Fenoprofen[37]	Oxaprozin
Ibuprofen [37, 39, 54, 55]	Allopurinol [56]	Cyclosporine	Ketoprofen [37]	Piroxicam [37]
Lisinopril	Candesartan	Diflunisal	Ketorolac [54]	Telmisartan
Meloxicam [38, 39, 54]	Capreomycin	Enalaprilat	Mefenamate	
Negative controls				
Benzonatate	Clozapine	Flavoxate	Methocarbamol	Ramelteon
Ketoconazole	Cosyntropin	Flutamide	Miconazole	Rizatriptan
Loratadine	Dacarbazine	Frovatriptan	Nelfinavir	Scopolamine
Metaxalone	Darbepoetin alfa	Gatifloxacin	Neostigmine	Simethicone
Temazepam	Darifenacin	Griseofulvin	Nortriptyline	Sodium phosphate, monobasic
Acarbose	Darunavir	Hyoscyamine	Orlistat	Tetrahydrocannabinol
Adenosine	Dicyclomine	Imipramine	Paromomycin	Thiabendazole
Almotriptan	Disulfiram	Infliximab	Penicillin V	Thiothixene
Amylases	Eletriptan	Ketotifen	Phentermine	Tinidazole
Benzocaine	Endopeptidases	Lactulose	Phentolamine	Urea
Bromfenac	Entecavir	Lipase	Prilocaine	Vitamin A
Chlorambucil	Ergotamine	Mebendazole	Primidone	Zafirlukast
Chlorazepate	Ferrous gluconate	Methenamine	Prochlorperazine	
Acute liver injury				
Positive controls				
Allopurinol [33]	**Niacin**	Alatrofloxacin	Imatinib	Stavudine
Carbamazepine [31]	**Nifedipine**	Bortezomib	Infliximab	Sulfisoxazole
Celecoxib	**Nitrofurantoin**	Bosentan	Interferon beta-1a	Tenofovir
Ciprofloxacin	**Nortriptyline**	Busulfan	Isoniazid [33]	Thiabendazole
Cyclosporine	**Ofloxacin**	Captopril [33]	Itraconazole [57]	Thioguanine
Diltiazem	**Oxaprozin**	Caspofungin	Lamivudine	Tipranavir
Erythromycin [31–33]	**Pioglitazone**	Clozapine	Methimazole	Tolcapone
Etodolac	**Piroxicam**	Dacarbazine	Methyldopa	Tolmetin
Fluconazole [58]	**Quinapril**	Darunavir	Moexipril	Trovafloxacin
Ibuprofen	**Ramipril**	Didanosine	Nefazodone	Voriconazole
Indomethacin [33]	**Sulindac**	Disulfiram	Nevirapine[59]	Zafirlukast
Ketorolac	**Tamoxifen** [60]	Efavirenz	Norfloxacin	Zalcitabine
Lamotrigine	**Terbinafine** [57]	Enalaprilat [33]	Orlistat	Zidovudine
Levofloxacin	**Trandolapril**	Felbamate	Penicillamine	

Table 1 continued

Lisinopril	**Valproate** [31]	Flutamide	Posaconazole	
Methotrexate	Acetazolamide	Gemcitabine	Propylthiouracil	
Naproxen	Abacavir	Gemifloxacin	Rifampin	
Negative controls				
Adenosine	**Lactulose**	Almotriptan	Ketotifen	Primidone
Benzocaine	**Miconazole**	Amylases	Lipase	Propantheline
Benzonatate	**Oxybutynin**	Cosyntropin	Lithium citrate	Sodium phosphate, monobasic
Dicyclomine	**Penicillin V**	Droperidol	Methenamine	Tetrahydrocannabinol
Fluticasone	**Salmeterol**	Endopeptidases	Neostigmine	Tinidazole
Gatifloxacin	**Scopolamine**	Ergotamine	Paromomycin	
Griseofulvin	**Sitagliptin**	Ferrous gluconate	Phentermine	
Hyoscyamine	**Sucralfate**	Flavoxate	Phentolamine	
Acute myocardial infarction				
Positive controls				
Amlodipine	**Ketorolac** [61]	Almotriptan	Factor VIIa	Rizatriptan
Darbepoetin alfa	**Nabumetone** [36, 61]	Amoxapine	Fenoprofen [61]	Salsalate
Dipyridamole	**Nifedipine**	Bromocriptine	Flurbiprofen [61]	Tolmetin [61]
Epoetin Alfa	**Nortriptyline**	Desipramine	Frovatriptan	Zolmitriptan
Estradiol	**Oxaprozin** [61]	Diflunisal	Imipramine	
Estrogens, conjugated (USP) [62]	**Piroxicam** [36, 61, 63]	Eletriptan	Ketoprofen [36, 61]	
Etodolac [36, 64]	**Sulindac** [61, 63]	Enalaprilat	Moexipril	
Indomethacin [34–36, 61]	**Sumatriptan**	Estropipate	Naratriptan	
Negative controls				
Benzonatate	**Ramelteon**	Chlorothiazide	Methenamine	Stavudine
Clindamycin	**Salmeterol**	Cosyntropin	Methimazole	Sulfasalazine
Dicyclomine	**Scopolamine**	Darifenacin	Miconazole	Sulfisoxazole
Fluticasone	**Sitagliptin**	Didanosine	Nelfinavir	Tetrahydrocannabinol
Gatifloxacin	**Sucralfate**	Droperidol	Nevirapine	Thiabendazole
Hyoscyamine	**Temazepam**	Endopeptidases	Paromomycin	Thiothixene
Ketoconazole	**Terbinafine**	Entecavir	Pemoline	Tinidazole
Lactulose	**Urea**	Ferrous gluconate	Penicillamine	Tipranavir
Loratadine	Acarbose	Flavoxate	Posaconazole	Vitamin A
Metaxalone	Acetazolamide	Flutamide	Prilocaine	Zafirlukast
Methocarbamol	Amylases	Ketotifen	Primidone	
Penicillin V	Bromfenac	Lipase	Propantheline	
Prochlorperazine	Chlorambucil	Lithium citrate	Simethicone	
Oxybutynin	Chlorazepate	Mebendazole	Sodium phosphate, monobasic	

Table 1 continued

GI bleed				
Positive controls				
Citalopram	**Fluoxetine** [18, 40, 65–71]	**Nabumetone** [72, 73]	**Oxaprozin**	Mefenamate [74]
Clindamycin	**Ibuprofen** [73, 75–82]	**Naproxen** [72, 73, 75, 77–83]	Diflunisal	Sulindac
Clopidogrel [18, 84]	**Indomethacin** [73, 75, 77–81]	**Piroxicam** [73, 77–83]	Fenoprofen	Tolmetin
Escitalopram	**Ketorolac** [78–80, 83]	**Potassium chloride**	Flurbiprofen [73, 78]	Valdecoxib
Etodolac [73, 77]	**Meloxicam** [73, 75, 77–80, 85]	**Sertraline** [18, 40, 65–71]	Ketoprofen [73, 75, 78–81]	
Negative controls				
Adenosine	**Pioglitazone**	Bromfenac	Ketotifen	Prilocaine
Benzonatate	**Prochlorperazine**	Chlorambucil	Lamivudine	Propantheline
Dicyclomine	**Rosiglitazone**	Chlorazepate	Lipase	Simethicone
Epoetin alfa	**Salmeterol**	Cosyntropin	Lithium citrate	Stavudine
Fluticasone	**Scopolamine**	Dacarbazine	Mebendazole	Tetrahydrocannabinol
Hyoscyamine	**Sitagliptin**	Darifenacin	Miconazole	Thiabendazole
Ketoconazole	**Sucralfate**	Disulfiram	Moexipril	Thiothixene
Lactulose	**Temazepam**	Droperidol	Neostigmine	Tinidazole
Loratadine	**Terbinafine**	Endopeptidases	Nevirapine	Vitamin A
Metaxalone	**Urea**	Entecavir	Orlistat	Zidovudine
Methocarbamol	Abacavir	Ergotamine	Paromomycin	
Nitrofurantoin	Acarbose	Ferrous gluconate	Pemoline	
Oxybutynin	Amylases	Griseofulvin	Phentermine	
Penicillin V	Benzocaine	Itraconazole	Phentolamine	

Supplemental (MSLR, 1.2 m persons), MarketScan® Medicare Supplemental Beneficiaries (MDCR, 4.6 m persons), MarketScan® Multi-State Medicaid (MDCD, 10.8 m persons), MarketScan® Commercial Claims and Encounters (CCAE, 46.5 m persons), and the General Electric Centricity™ (GE, 11.2 m persons) database. GE is an electronic health record (EHR) database, the other four databases contain administrative claims data. Additional detail about each database is available elsewhere in this Supplement [26]. This criterion was imposed simply to ensure the drug was observable in our databases, that is, they had an associated NDC and/or HCPCS code that was current with the reference period for this analysis. All medical products that were vaccines or were only administered as ophthalmologic or topical routes of administration were excluded, along with positive and negative control drugs for which we could not identify a comparator drug with the same indication, as defined algorithmically using the OMOP vocabulary [27].

Once the reference set was constructed, we computed for each test case and every database the minimum detectable relative risk (MDRR), that is the smallest relative risk that can be detected at an alpha of 0.05 and requiring statistical power of 0.80 [28], and determined the number of controls with MDRR $\leq$1.25. This analysis was performed to assess how many of the test cases within each outcome would have sufficient sample size with adequate power in each database to be able to be studied in our methodological evaluations. MDRR $\leq$1.25 was selected as a fairly stringent threshold on the basis that the recent product withdrawal of rosiglitazone was informed by a meta-analysis whose point estimate for the risk of myocardial infarction was OR = 1.28 [29]. The prevalence of each outcome within each database is presented as context for evaluating the MDRR threshold, based on outcome definitions described elsewhere [30].

3 Results

Table 1 provides the list of all test cases, with associated published evidence. In total, we identified 165 positive controls and 234 negative controls. A complete listing of the test cases is available in Appendix 1 in ESM. In addition to providing the drug-outcome pair with ground truth classification of 'positive' or 'negative', Appendix 1 in ESM includes the URL to the structured product label that contained warning for positive controls, the level of evidence assigned by Tisdale et al. [25], attributes about the drug [including Anatomical Therapeutic Chemical (ATC) classification system, primary indication, and primary comparator drug used when evaluating the new user cohort design], and for each database, provides the number of exposed patients, average length of exposure, and minimum detectable relative risk.

The total number of positive controls and negative controls per outcome is shown in Fig. 2 in the leftmost bars. For every outcome more than 80 test cases were identified, with at least 20 positive and negative controls. For all outcomes except acute liver injury, there were more negative controls than positive controls. For each database, Fig. 2 also shows the number of test cases available for use in evaluations after imposing the restriction that the minimum detectable relative risk $\leq$1.25. A large sample of test cases was preserved for most database-outcome scenarios, although when studying acute renal failure in MSLR and GE we are limited to fewer than 10 positive and negative controls, thereby increasing the uncertainty of the evaluation metrics. Table 2 provides prevalence of each of the four outcomes across the 5 database and shows that the acute renal failure is the least commonly occurring outcome.

Table 3 shows the level of evidence identified by Tisdale. While acute liver injury had the largest total number of positive controls, it had the fewest number of test cases supported by evidence from randomized trials. Sixty-nine percent of test cases for acute liver injury were only supported by case reports or case series. In contrast, the majority of test cases for acute kidney injury (13/24) and upper GI bleeding (15/24) were supported by randomized trials. This is largely due to the prevalence of trial evidence around nonsteroidal anti-inflammatory drugs (NSAIDs), which account for many of the test cases for these two outcomes.

Figure 3 highlights the range of effect estimates observed from publications identified for a test case in each outcome. The full list of estimates extracted from the literature is available in Appendix 2 in ESM. For erythromycin and its relation to acute liver injury, we identified three published studies; de Abajo et al. [31] and Carson et al. [32] provided estimates from case-control studies that were largely consistent, but with wide confidence intervals suggesting the effect could only be bound between RR = 1.5 and RR = 20. Sabate et al. [33] published a cohort study that suggested the effect size could be even larger, but also with substantial uncertainty around the estimate. Three studies of the effect of indomethacin and acute myocardial infarction suggest the effect estimate is likely less than RR = 2 [34–36]. Three case-control studies that assessed the effect of naproxen on acute kidney injury showed substantial heterogeneity, with Griffen et al. [37], suggesting essentially no effect, Schneider et al. [38] highlighting a significant effect with RR <3, and Huerta et al. [39] producing a non-significant estimated effect with RR >3. Multiple studies highlighted a positive and statistically significant association between sertraline and upper GI bleeding but a published meta-analysis could only bound the purported effect between RR = 1.44 and RR = 3.85 [40].

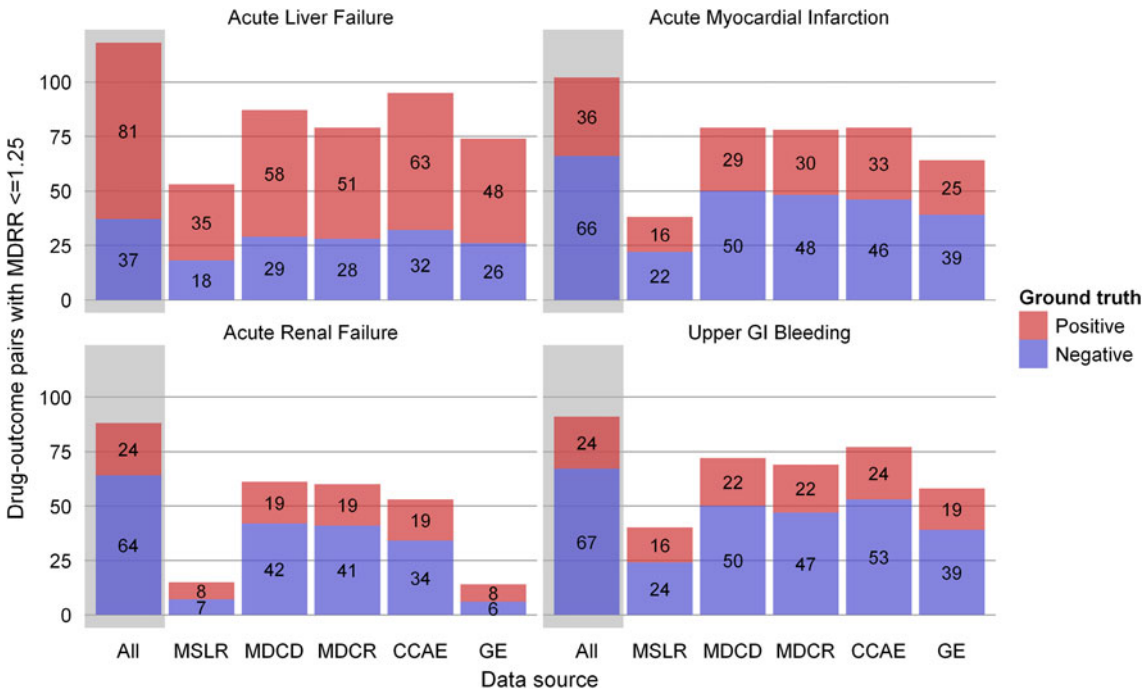

Fig. 2 Number of positive and negative controls. *MDRR* minimum detectable relative risk, *MSLR* MarketScan Lab Supplemental, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE centricity

Table 2 Prevalence of four outcomes across five databases

Condition group name	CCAE		MSLR		MDCD		MDCR		GE	
	Persons with event	Prev (%)	Persons with event	Prev (%)	Persons with event	Prev (%)	Persons with event	Prev (%)	Persons with event	Prev (%)
Acute kidney injury	114,472	0.2	7,403	0.6	97,910	0.9	201,769	4.4	12,553	0.1
Acute liver injury	1,235,711	2.7	63,505	5.2	274,723	2.5	264,122	5.8	186,677	1.7
Acute myocardial infarction	731,792	1.6	39,767	3.2	221,926	2.1	665,396	14.6	140,335	1.3
Upper GI bleeding	782,127	1.7	32,935	2.7	206,906	1.9	342,937	7.5	108,882	1.0

Prev prevalence, *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 Level of evidence for positive control test cases

Tisdale level of evidence	Acute liver injury	Acute kidney injury	Acute myocardial Infarction	Upper gastrointestinal bleeding	Total
A: evidence from one or more randomized controlled clinical trials	6	13	16	18	53
B: support from nonrandomized clinical trials, prospective observational studies, cohort studies, retrospective studies, case-control studies, meta-analyses, and/or postmarketing surveillance studies	19	8	0	5	32
C: evidence from one or more published case reports or case series	56	3	20	1	80
Total	81	24	36	24	165

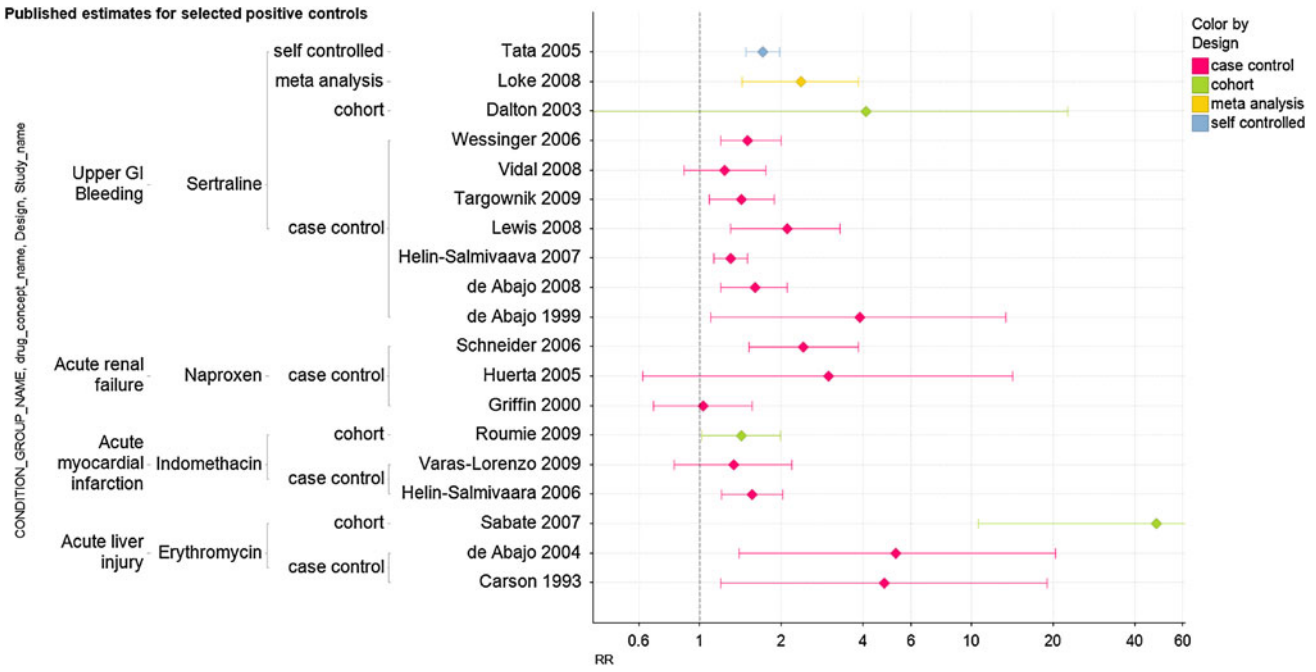

Fig. 3 Published estimates from selected positive controls

4 Discussion

Proper evaluation of drug safety questions requires scientific evidence to form the basis for medical and regulatory decision-making. In order to interpret emerging evidence about safety concerns where beliefs about the potential causal effect are not fully established, it is important to first understand the expected accuracy of the information source. Most researchers have a qualitative notion of a hierarchy of evidence, with randomized trials representing the highest quality of information, case reports providing the weakest objective evidence, and population-based observational studies falling somewhere in between. However, there is no formal quantitative assessment of the hierarchy of evidence that provides clear empirical guidance on the degree to which complementary data sources can accurately inform decision-making. Such a quantitative summary would require measuring the performance of study designs by comparing estimated effects with some pre-defined 'gold standard' benchmark.

This paper describes the steps taken in the OMOP research program to establish a reference set of positive controls and negative controls that can serve as that benchmark. Our work complements the efforts of others to construct reference sets to facilitate methodological research, but is unique in attempting to provide a comprehensive positive control set for specific outcomes of interest in conjunction with a large set of negative controls. The 399 test cases developed in this effort represent the largest known reference set for evaluating methods development for studying. This large reference set should enable broad evaluation across outcomes, as well as in-depth investigations of performance within specific events, and should yield additional insights about specific circumstances of interest.

While our motivating objective was comprehensiveness, we applied highly restrictive criteria to selecting positive controls and negative controls in an attempt to minimize the risk of misclassification amongst the test cases. Misclassification may still exist, as evidence supporting the assertion of a positive control could be incorrect. It has been shown that product labeling can be inconsistent across manufacturers within the same product [41]. Moreover, while we eliminated all test cases where we observed any conflicting evidence in the literature that was contrary to the product label, it is possible that test cases that had consistent support for a positive effect was the manifestation of a common bias across the evidence sources. We did not independently grade the quality of evidence available for all test cases, instead relying on the systematic review previously conducted by Tisdale. Given that the motivation for creating the reference set is to study the performance characteristics of observational studies which are not yet fully understood, it is possible that the conflicting evidence we identified is based on poor quality studies which are unnecessarily limiting our sample of positive and negative controls. Moreover, absence of evidence for a negative control cannot be interpreted as definitive evidence of

absence. The wide evidentiary separation using this 'all or nothing' approach to selecting positive and negative controls may in fact result in methods evaluations that are optimistically biased, since the test cases do not include as many drug-outcome pairs with smaller effects or those which are more ambiguous with conflicting evidence from multiple sources.

In the ideal world, a reference set for drug safety would not only provide a binary indication of positive effect or no effect, but would also provide a measure of the true effect size. Having a true effect size would broaden the available approaches that could be taken to evaluate methods performance, where methods estimate the effect through measures of association. With a binary classification of ground truth, methods evaluation is limited to measures of discrimination, such as area under receiver operating characteristic curve [42]. With a true effect size, methods evaluation can be expanded to include measures of accuracy, such as bias and mean squared error, and calibration, such as coverage probability. The results of our literature review illustrate that for the positive controls, while we may safely assume that the true effect is positive, we cannot assign a specific 'true RR' value. The literature shows how the variability in the estimates for the same positive controls, and how much uncertainty exists around each estimate. For negative controls, because there is no evidence suggesting an effect, we can only assume the true RR = 1 with the acknowledgment we cannot validate the accuracy of this assumption. These practical challenges in defining a reference set highlight the value of complementing methods evaluations in real-world databases with studies in simulated dataset where reference sets can be artificially constructed with known effect sizes.

Underlying the use of a reference set for methods evaluation is the assumption that negative controls are exchangeable with positive controls. When measuring a method's ability to distinguish between positive and negative controls, we want the only differentiation between the test cases to be their true effect size, much in the same way that we measure the efficacy of a treatment based on randomization of patients such that the only differentiation between the treatment groups is the exposure status. In practice, we observe asymmetry between positive and negative controls. For example, drugs identified as positive controls tend to be more commonly used, have a longer average duration of exposure, and are used to treat more serious diseases than the drugs identified as negative controls. This imbalance is due in part to the non-random process used to select test cases, as evidence to support positive controls is more likely to be identified among prevalent drugs with large public health impact which are more likely to be studied by the broader research community. It also can be attributed to the inherent benefit-risk tradeoffs associated with medical treatments; drugs that are suspected to cause serious harms—such as myocardial infarction, renal impairment, or hepatic damage—need to demonstrate sufficient clinical benefits for important indicated diseases to offset those risks, while treatments for less severe diseases are more likely to be well tolerated with fewer side effects. Any of these differences between positive and negative controls may influence the degree of confounding differentially observed among the test cases, and could potentially bias the measure of discrimination used in methods evaluation. In methodological experiments, one way to attenuate this potential bias is to restrict the test cases to those pairs with sufficient sample size. In Fig. 2, we highlighted the database-specific consequences of limiting test case to those with MDRR $\leq$1.25. This threshold should not imply that effect estimates of that magnitude should be necessarily deemed as credible, but merely that sample size should not present a major consideration or limitation when assessing the performance of methods. Eliminating the potential source of error due to insufficient sample within each drug-outcome pair comes with the reduced precision of the performance measurement, due to smaller sample of test cases used in the evaluation. Another limitation of restricting test cases on MDRR is that the results of a methodological evaluation may not be as generalizable to situations where drugs are less commonly prescribed or for newly marketed products where exposures have not yet accumulated. Other thresholds, such as MDRR $\leq$2.0 may be considered as some form of compromise in this regard, but may result in observing lower predictive accuracy due to data size limitations. Coloma et al. [43] and Reich et al. [44] provide thoughtful discussion about why sample size requires further consideration when developing an risk identification system.

The utility of evidence from methods evaluation using a reference set rests with the belief that the drug-outcome scenarios represented by the sample of positive and negative controls used in retrospective evaluation can be generalized to the future drug-outcome association of interest. This reference set is limited to its focus on four outcomes. It can be used to study the performance within these outcomes, but results may not generalize to other outcomes of interest. Since becoming available, this reference set has been applied to study the performance of alternative methods for risk identification in observational healthcare data [45–52] as well as in a systematic evaluation of disproportionality analysis methods applied to a spontaneous adverse advent reporting database by Harpaz et al. [53]. In both bodies of work, substantial differences in methodological performance was observed across outcomes. This may not be unexpected given the inherent differences in the

biological and clinical presentations of the outcomes themselves and the differential challenges in identify drug-induced events through the medical product development lifecycle. It may also point to inconsistent confidence in the classification of test cases across the outcomes based on the variable quantity and quality of evidence available for defining positive and negative controls. Despite these limitations, the process followed to generate the test cases for these four outcomes may be able to be replicated to additional outcomes of interest. For example, Tisdale et al. [25], offers systematic literature reviews for 11 additional events of importance: convulsions, extrapyramidal disorders, depression, pancreatitis, thrombocytopenia, neutropenia, aplastic anemia, venous thromboembolism, anaphylaxis, peripheral neuropathy, and Stevens Johnson Syndrome. The use of NLP to extract adverse events from structure product label could easily be extended to other outcomes once definitions were established. Thus, the reference set provided could be regarded as the starting point within an evolving landscape to expand our evidence base for drug safety research.

5 Conclusions

Establishing a comprehensive reference set, across a large array of health outcomes, would be a tremendous public utility that could provide a strong foundation and stimulate expanded research and development within the methodology community for years to come.

Acknowledgments The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health (FNIH) through generous contributions from the following: Abbott, Amgen Inc., AstraZeneca, Bayer Healthcare Pharmaceuticals, Inc., Biogen Idec, Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Janssen Research and Development, Lundbeck, Inc., Merck & Co., Inc., Novartis Pharmaceuticals Corporation, Pfizer Inc., Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi-aventis, Schering-Plough Corporation, and Takeda. Drs. Ryan and Schuemie are employees of Janssen Research and Development. Dr. Schuemie received a fellowship from the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration. Drs. Duke, Schuemie and Hartzema have previously received funding from FNIH. Emily Welebob and Sarah Valentine have no conflicts of interest to declare.

This article was published in a supplement sponsored by the Foundation for the National Institutes of Health (FNIH). The supplement was guest edited by Stephen J.W. Evans. It was peer reviewed by Olaf H. Klungel who received 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. FDA. Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment. US FDA Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research; 2005.
2. Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965;58:295–300.
3. FDA. The Sentinel Initiative: A National Strategy for Monitoring Medical Product Safety. May 2008 [cited 2012 September 15]. http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm089474.htm.
4. Lindquist M, Ståhl M, Bate A, Edwards IR, Meyboom RH. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. Drug Safety. 2000;23(6):533–42.
5. Hauben M, Reich L. Safety related drug-labelling changes: findings from two data mining algorithms. Drug Safety. 2004;27(10):735–44.
6. Hochberg AM, Hauben M, Pearson RK, O'Hara DJ, Reisinger SJ, Goldsmith DI, et al. An evaluation of three signal-detection algorithms using a highly inclusive reference event database. Drug Safety. 2009;32(6):509–25.
7. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol. 2003;158(9):915–20.
8. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009;20(4):512–22.
9. Woodward M. Epidemiology study design and data analysis. London: Chapman & Hall/CRC; 1999.
10. Whitaker H. The self controlled case series method. BMJ 2008;337:a1069. http://dx.doi.org/10.1136/bmj.a1069.
11. Norén N, Hopstadius J, Bate A, Star K, Edwards IR. Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Discov. 2010;20(3):361–87.
12. Schuemie MJ. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. Pharmacoepidemiol Drug Saf. 2011;20(3):292–9.
13. Ryan PB, Powell GE, Pattishall EN, Beach KJ. Performance of screening multiple observational databases for active drug safety surveillance. Poster presented at the 25 annual meeting of the International Society of Pharmacoepidemiology, Providence, Rhode Island, 16–19 August 2009.
14. 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.
15. 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 Safety. 2013;36(1):13–23.
16. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010;153(9):600–6.
17. 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.
18. Wessinger S, Kaplan M, Choi L, Williams M, Lau C, Sharp L, et al. Increased use of selective serotonin reuptake inhibitors in patients admitted with gastrointestinal haemorrhage: a multicentre retrospective analysis. Aliment Pharmacol Ther. 2006;23(7):937–44.

19. Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salame G, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiol Drug Safety. 2009; 18(12):1176–84.
20. Katz AJ, Ryan PB, Racoosin JA, Stang PE. Assessment of case definitions for identifying acute liver injury in large observational databases. Drug Safety. 2013;36(8):651–61.
21. Duke J, Friedlin J, Ryan P. A quantitative analysis of adverse events and "overwarning" in drug labeling. Arch Intern Med. 2011;171(10):944–6.
22. Duke JD, Friedlin J. ADESSA: a real-time decision support service for delivery of semantically coded adverse drug event data. AMIA Annu Symp Proc. 2010;2010:177–81.
23. Friedlin J, Duke J. Applying natural language processing to extract codify adverse drug reaction in medication labels. 2010 [cited 2013 January 3]. http://omop.org/sites/default/files/omop_white_paper_friedlin_08_26_10.pdf.
24. Friedlin J, Duke J. Exploration of four outcomes: outcomes and labeling information, in conjunction with other evidence May 2011 [cited 2013 January 3]. http://omop.org/sites/default/files/OMOP%20Report_Duke_Friedlin_05_16_11%20Exploration%20of%20Four%20Outcomes.pdf.
25. Tisdale J, Miller D. Drug-induced diseases: prevention, detection, and management. 2nd ed. USA: American Society of Health-System Pharmacists; 2010.
26. 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.
27. Ryan PB, Madigan D. Selecting comparators in active surveillance analyses. 2010 [cited 2013 January 3]. http://omop.org/sites/default/files/OMOP%20-%20Selecting%20comparators%20in%20active%20surveillance%20analyses.pdf.
28. 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.
29. Nissen SE, Wolski K. Rosiglitazone revisited: an updated meta-analysis of risk for myocardial infarction and cardiovascular mortality. Arch Intern Med. 2010;170(14):1191–201.
30. Hansen RA, Gray MD, Fox BI, Hollingsworth JC, Gao J, Zeng P. How well do various health outcome definitions used in observational studies identify cases that are consistent with expert opinion? Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0104-0.
31. de Abajo FJ, Montero D, Madurga M, García Rodríguez LA. Acute and clinically relevant drug-induced liver injury: a population based case-control study. Br J Clin Pharmacol. 2004;58(1):71–80.
32. Carson JL, Strom BL, Duff A, Gupta A, Shaw M, Lundin FE, et al. Acute liver disease associated with erythromycins, sulfonamides, and tetracyclines. Ann Intern Med. 1993;119(1):576–83.
33. Sabate M, Ibanez L, Perez E, Vidal X, Buti M, Xiol X, et al. Risk of acute liver injury associated with the use of drugs: a multicentre population survey. Aliment Pharmacol Ther. 2007; 25(12):1401–9.
34. Roumie CL, Choma NN, Kaltenbach L, Mitchel EF Jr, Arbogast PG, Griffin MR. Non-aspirin NSAIDs, cyclooxygenase-2 inhibitors and risk for cardiovascular events-stroke, acute myocardial infarction, and death from coronary heart disease. Pharmacoepidemiol Drug Saf. 2009;18(11):1053–63.
35. Varas-Lorenzo C, Castellsague J, Stang MR, Perez-Gutthann S, Aguado J, Rodriguez LA. The use of selective cyclooxygenase-2 inhibitors and the risk of acute myocardial infarction in Saskatchewan, Canada. Pharmacoepidemiol Drug Safety. 2009; 18(11):1016–25.
36. Helin-Salmivaara A, Virtanen A, Vesalainen R, Gronroos JM, Klaukka T, Idanpaan-Heikkila JE, et al. NSAID use and the risk of hospitalization for first myocardial infarction in the general population: a nationwide case-control study from Finland. Eur Heart J. 2006;27(14):1657–63.
37. Griffin MR, Yared A, Ray WA. Nonsteroidal antiinflammatory drugs and acute renal failure in elderly persons. Am J Epidemiol. 2000;151(5):488–96.
38. Schneider V, Levesque LE, Zhang B, Hutchinson T, Brophy JM. Association of selective and conventional nonsteroidal antiinflammatory drugs with acute renal failure: a population-based, nested case-control analysis. Am J Epidemiol. 2006; 164(9):881–9.
39. Huerta C, Castellsague J, Varas-Lorenzo C, García Rodríguez LA. Nonsteroidal anti-inflammatory drugs and risk of ARF in the general population. Am J Kidney Dis. 2005;45(3):531–9.
40. Loke YK, Trivedi AN, Singh S. Meta-analysis: gastrointestinal bleeding due to interaction between selective serotonin uptake inhibitors and non-steroidal anti-inflammatory drugs. Aliment Pharmacol Ther. 2008;27(1):31–40.
41. Duke J, Friedlin J, Li X. Consistency in the safety labeling of bioequivalent medications. Pharmacoepidemiol Drug Saf. 2013;22:294–301.
42. Cantor SB, Kattan MW. Determining the area under the ROC curve for a binary diagnostic test. Med Decis Making. 2000; 20(4):468–70.
43. Coloma PM, Trifirò G, Schuemie MJ, Gini R, Herings R, Hippisley-Cox J, et al. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? Pharmacoepidemiol Drug Safety. 2012;21(6):611–21.
44. Reich CG, Ryan PB, Suchard MA. The impact of drug and outcome prevalence on the feasibility and performance of analytical methods for a risk identification and analysis system. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0112-0.
45. Schuemie MJ, Madigan D, Ryan PB. Empirical performance of longitudinal gamma poisson shrinker (LGPS) and longitudinal evaluation of observational profiles of adverse events related to drugs (LEOPARD): lessons for developing a risk identification and analysis system. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0107-x.
46. Norén GN, Bergvall T, Ryan PB, Juhlin K, Schuemie MJ, Madigna 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.
47. 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.
48. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0099-6.
49. 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.
50. 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.
51. DuMouchel B, Ryan PB, Schuemie MJ, Madigan D. Evaluation of disproportionality safety signaling applied to health care

databases. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0106-y.
52. Schuemie MJ, Gini R, Coloma PM, Straatman H, Herings RMC, Pedersen L, et al. Replication of the OMOP experiment in Europe: evaluating methods for risk identification in electronic health record databases. Drug Saf. 2013 (in this supplement issue). doi:10.1007/s40264-013-0109-8.
53. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther. 2012;91(6):1010–21.
54. Winkelmayer WC, Waikar SS, Mogun H, Solomon DH. Nonselective and cyclooxygenase-2-selective NSAIDs and acute kidney injury. Am J Med. 2008;121(12):1092–8.
55. Murray MD, Brater DC, Tierney WM, Hui SL, McDonald CJ. Ibuprofen-associated renal impairment in a large general internal medicine practice. Am J Med Sci. 1990;299(4):222–9.
56. Hung CC, Liu WC, Kuo MC, Lee CH, Hwang SJ, Chen HC. Acute renal failure and its risk factors in Stevens–Johnson syndrome and toxic epidermal necrolysis. Am J Nephrol. 2009;29(6):633–8.
57. García Rodríguez LA, Duque A, Castellsague J, Perez-Gutthann S, Stricker BH. A cohort study on the risk of acute liver injury among users of ketoconazole and other antifungal drugs. Br J Clin Pharmacol. 1999;48(6):847–52.
58. Fischer MA, Winkelmayer WC, Rubin RH, Avorn J. The hepatotoxicity of antifungal medications in bone marrow transplant recipients. Clin Infect Dis. 2005;41(3):301–7.
59. Ouyang DW, Shapiro DE, Lu M, Brogly SB, French AL, Leighty RM, et al. Increased risk of hepatotoxicity in HIV-infected pregnant women receiving antiretroviral therapy independent of nevirapine exposure. AIDS. 2009;23(18):2425–30.
60. Bruno S, Maisonneuve P, Castellana P, Rotmensz N, Rossi S, Maggioni M, et al. Incidence and risk factors for non-alcoholic steatohepatitis: prospective study of 5408 women enrolled in Italian tamoxifen chemoprevention trial. BMJ. 2005;330(7497):932.
61. Solomon DH, Avorn J, Stürmer T, Glynn RJ, Mogun H, Schneeweiss S. Cardiovascular outcomes in new users of coxibs and nonsteroidal antiinflammatory drugs: high-risk subgroups and time course of risk. Arthritis Rheum. 2006;54(5):1378–89.
62. Khader YS, Rice J, John L, Abueita O. Oral contraceptives use and the risk of myocardial infarction: a meta-analysis. Contraception. 2003;68(1):11–7.
63. Mangoni AA, Woodman RJ, Gaganis P, Gilbert AL, Knights KM. Use of non-steroidal anti-inflammatory drugs and risk of incident myocardial infarction and heart failure, and all-cause mortality in the Australian veteran community. Br J Clin Pharmacol. 2010;69(6):689–700.
64. Warner JJ, Weideman RA, Kelly KC, Brilakis ES, Banerjee S, Cunningham F, et al. The risk of acute myocardial infarction with etodolac is not increased compared to naproxen: a historical cohort analysis of a generic COX-2 selective inhibitor. J Cardiovasc Pharmacol Ther. 2008;13(4):252–60.
65. Dalton SO, Johansen C, Mellemkjaer L, Norgard B, Sorensen HT, Olsen JH. Use of selective serotonin reuptake inhibitors and risk of upper gastrointestinal tract bleeding: a population-based cohort study. Arch Intern Med. 2003;163(1):59–64.
66. de Abajo FJ, García-Rodríguez LA. Risk of upper gastrointestinal tract bleeding associated with selective serotonin reuptake inhibitors and venlafaxine therapy: interaction with nonsteroidal anti-inflammatory drugs and effect of acid-suppressing agents. Arch Gen Psychiatry. 2008;65(7):795–803.
67. de Abajo FJ, Rodriguez LA, Montero D. Association between selective serotonin reuptake inhibitors and upper gastrointestinal bleeding: population based case-control study. BMJ. 1999;319(7217):1106–9.
68. Helin-Salmivaara A, Huttunen T, Gronroos JM, Klaukka T, Huupponen R. Risk of serious upper gastrointestinal events with concurrent use of NSAIDs and SSRIs: a case-control study in the general population. Eur J Clin Pharmacol. 2007;63(4):403–8.
69. Lewis JD, Strom BL, Localio AR, Metz DC, Farrar JT, Weinrieb RM, et al. Moderate and high affinity serotonin reuptake inhibitors increase the risk of upper gastrointestinal toxicity. Pharmacoepidemiol Drug Saf. 2008;17(4):328–35.
70. Targownik LE, Bolton JM, Metge CJ, Leung S, Sareen J. Selective serotonin reuptake inhibitors are associated with a modest increase in the risk of upper gastrointestinal bleeding. Am J Gastroenterol. 2009;104(6):1475–82.
71. Vidal X, Ibanez L, Vendrell L, Conforti A, Laporte LR, Spanish-Italian Collaborative Group for the Epidemiology of Gastrointestinal B. Risk of upper gastrointestinal bleeding and the degree of serotonin reuptake inhibition by antidepressants: a case-control study. Drug Safety. 2008;31(2):159–68.
72. 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.
73. García Rodríguez LA, Hernandez-Diaz S. Relative risk of upper gastrointestinal complications among users of acetaminophen and nonsteroidal anti-inflammatory drugs. Epidemiology. 2001;12(5):570–6.
74. de Abajo FJ, García Rodríguez LA. Risk of upper gastrointestinal bleeding and perforation associated with low-dose aspirin as plain and enteric-coated formulations. BMC Clin Pharmacol. 2001;1:1.
75. García Rodríguez LA, Barreales Tolosa L. Risk of upper gastrointestinal complications among users of traditional NSAIDs and COXIBs in the general population. Gastroenterology. 2007;132(2):498–506.
76. Grimaldi-Bensouda L, Abenhaim L, Michaud L, Mouterde O, Jonville-Bera AP, Giraudeau B, et al. Clinical features and risk factors for upper gastrointestinal bleeding in children: a case-crossover study. Eur J Clin Pharmacol. 2010;66(8):831–7.
77. Hippisley-Cox J, Coupland C, Logan R. Risk of adverse gastrointestinal outcomes in patients taking cyclo-oxygenase-2 inhibitors or conventional non-steroidal anti-inflammatory drugs: population based nested case-control analysis. BMJ. 2005;331(7528):1310–6.
78. Lanas A, García-Rodríguez LA, Arroyo MT, Gomollon F, Feu F, Gonzalez-Perez A, et al. Risk of upper gastrointestinal ulcer bleeding associated with selective cyclo-oxygenase-2 inhibitors, traditional non-aspirin non-steroidal anti-inflammatory drugs, aspirin and combinations. Gut. 2006;55(12):1731–8.
79. Laporte JR, Ibanez L, Vidal X, Vendrell L, Leone R. Upper gastrointestinal bleeding associated with the use of NSAIDs: newer versus older agents. Drug Safety. 2004;27(6):411–20.
80. Massó Gonzalez EL, Patrignani P, Tacconelli S, García Rodríguez LA. Variability among nonsteroidal antiinflammatory drugs in risk of upper gastrointestinal bleeding. Arthritis Rheum. 2010;62(6):1592–601.
81. Mellemkjaer L, Blot WJ, Sorensen HT, Thomassen L, McLaughlin JK, Nielsen GL, et al. Upper gastrointestinal bleeding among users of NSAIDs: a population-based cohort study in Denmark. Br J Clin Pharmacol. 2002;53(2):173–81.
82. Rahme E, Nedjar H. Risks and benefits of COX-2 inhibitors vs non-selective NSAIDs: does their cardiovascular risk exceed their gastrointestinal benefit? A retrospective cohort study. Rheumatology. 2007;46(3):435–8.

83. Lanas A, Serrano P, Bajador E, Fuentes J, Sainz R. Risk of upper gastrointestinal bleeding associated with non-aspirin cardiovascular drugs, analgesics and nonsteroidal anti-inflammatory drugs. Eur J Gastroenterol Hepatol. 2003;15(2):173–8.
84. Opatrny L, Delaney JA, Suissa S. Gastro-intestinal haemorrhage risks of selective serotonin receptor antagonist therapy: a new look. Br J Clin Pharmacol. 2008;66(1):76–81.
85. Levesque LE, Brophy JM, Zhang B. The risk for myocardial infarction with cyclooxygenase-2 inhibitors: a population study of elderly adults. Ann Intern Med. 2005;142(7):481–9.