Drug Saf (2013) 36 (Suppl 1):S5–S14
DOI 10.1007/s40264-013-0102-2

COMMENTARY

Desideratum for Evidence Based Epidemiology

J. Marc Overhage · Patrick B. Ryan · Martijn J. Schuemie · Paul E. Stang

Abstract

Background There is great variation in choices of method and specific analytical details in epidemiological studies, resulting in widely varying results even when studying the same drug and outcome in the same database. Not only does this variation undermine the credibility of the research but it limits our ability to improve the methods.

Methods In order to evaluate the performance of methods and analysis choices we used standard references and a literature review to identify 164 positive controls (drug–outcome pairs believed to represent true adverse drug reactions), and 234 negative controls (drug–outcome pairs for which we have confidence there is no direct causal relationship). We tested 3,748 unique analyses (methods in combination with specific analysis choices) that represent the full range of approaches to adjusting for confounding in five large observational datasets on these controls. We also evaluated the impact of increasingly specific outcome definitions, and performed a replication study in six additional datasets. We characterized the performance of each method using the area under the receiver operator curve (AUC), bias, and coverage probability. In addition, we developed simulated datasets that closely matched the characteristics of the observational datasets into which we inserted data consistent with known drug–outcome relationships in order to measure the accuracy of estimates generated by the analyses.

Discussion We expect the results of this systematic, empirical evaluation of the performance of these analyses across a moderate range of outcomes and databases to provide important insights into the methods used in epidemiological studies and to increase the consistency with which methods are applied, thereby increasing the confidence in results and our ability to systematically improve our approaches.

1 Introduction

We do not often take the opportunity to study the "science of science": the methods and approaches that we use to

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.

J. M. Overhage
Siemens Health Services, Malvern, PA, USA

J. M. Overhage (✉)
Siemens Medical Solutions USA, Inc., 51 Valley Stream Parkway, MC B9K, Malvern, PA 19355, USA
e-mail: marc.overhage@siemens.com

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

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

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

advance our understanding. Much of the thinking in this domain has been carried out by philosophers rather than traditional scientists. Karl Popper, for example, wrote "…. I knew, of course, the most widely accepted answer to my problem: that science is distinguished from pseudo-science—or from 'metaphysics'—by its empirical method, which is essentially inductive, proceeding from observation or experiment" [1]. Essentially every scientist would agree with Popper's assertion and relies on data derived from observation or experiment to test their theories. Less common though, is for scientists to consider the next step which Popper described as "… distinguishing between a genuinely empirical method and a non-empirical or even a pseudo-empirical method—that is to say, a method which, although it appeals to observation and experiment, nevertheless does not come up to scientific standards. The latter method may be exemplified by astrology, with its stupendous mass of empirical evidence based on observation—on horoscopes and on biographies" [1]. In the authors' opinion, our current approach to observational research is not empirically based.

Popper formulated the notion of "problem of demarcation" to distinguish between scientific and pseudo-scientific theories and coined the term "falsifiability" to describe the test he developed to distinguish between scientific and pseudo-scientific theories. He defined falsifiability as the demonstration that a statement is false by finding a counterexample (an observation of the physical world that is incompatible with the statement). "The criterion of falsifiability is a solution to this problem of demarcation, for it says that statements or systems of statements, in order to be ranked as scientific, must be capable of conflicting with possible, or conceivable, observations." These observations should apply equally to the science of performing observational studies.

2 Epidemiological Studies Using Observational Data

Epidemiological studies using observational data have become an important tool in many areas including pharmacoepidemiology. A large proportion of assertions in the fields of medicine and public health (in addition to the fields of environment, climate change and psychology) are derived from observational studies: nearly 80,000 observational studies were published in the decade 1990–2000 and 263,557 in the following decade 2001–2011 [2]. In an observational study, researchers observe what happened during the course of events without intervening; they analyze the data and draw conclusions. Observational studies, when well designed, often yield effect estimates comparable to those of randomized controlled trials (RCT) [3–9]. However, observational studies have proven difficult to replicate. The percentage of observational studies whose results are not reproduced is estimated to range from 20 % [10] to 10 % [2]. Our limited ability to replicate these studies is worrisome and limits our confidence in their results. For example, in the late 1980s and early 1990s, observational studies relatively consistently suggested that hormone-replacement therapy (HRT) protected postmenopausal women against heart disease [11, 12] though there were some negative effects identified as well [13]. However, a large randomized clinical trial demonstrated increased risk of cardiovascular disease in incident users of HRT compared with nonusers [14]. Subsequent analysis of the observational studies focusing on incident HRT users did not find a reduction in risk of cardiovascular disease [15, 16].

More recently, separate studies of bisphosphonates and the risk of esophageal cancer using the same United Kingdom patient database came to different conclusions. One found no increase in patients' cancer risk while the second found a doubling of risk for developing cancer [17, 18]. Another example is two studies on risk of fracture associated with statins, both published in JAMA. The first study [19] found a statistically significant decreased risk, whereas the second study [20] found no effect. A later re-examination of the two studies attributed these differences to different methodological choices made by the researchers [21]. Ioannidis and Ioanna, in an analysis of published observational studies of 10 purported biomarkers for cardiovascular disease, found that many of the studies were undermined by publication bias and potential residual confounding and that they demonstrated little predictive ability despite the positive conclusions in the publications [22].

We might expect that replicating observational studies would be easier than physical experiments since the vagaries of experimental technique and potential for measurement error are less prominent—was the glassware adequately cleaned? Were the reagents prepared properly? Were instruments calibrated appropriately? Researchers in the field expect that "competent" laboratories do these things competently, but we also assume that the right things to do are well enough understood that they will be done routinely. If it turned out that exposing a compound to daylight rather than artificial light was a significant factor for example, it might be overlooked for some time. Until discovered, this factor might lead to variable results among even competent laboratories. Similarly, small procedural differences could result in variable results—one lab stores samples overnight on a refrigerator shelf with a temperature close to freezing and the other in a refrigerator door with a temperature several degrees higher which allows an enzymatic reaction to proceed. As anyone who has "brought up" a new analytical method can attest, these seemingly insignificant differences in technique or "tacit

knowledge" can be difficult to identify. Companies like Intel® have learned this lesson and go to great lengths to replicate even seemingly unimportant characteristics of a semi-conductor fabrication facility such as matching the length of a drain hose in a functioning facility exactly even if that length is longer than needed in the new facility. They have learned that failing to precisely replicate even the smallest details can result in differences in the process which are difficult to identify and have large commercial impact [23, 24].

Laboratory methods and equipment are often tested or verified using some form of reference standard. These might be carefully prepared samples from another laboratory, a carefully calibrated "reference" instrument or some other "gold standard". We use the instrument, following a clearly specified set of steps to measure the reference standard and "troubleshoot" the method until we produce results that match, within the limitations of the method, the expected results. Then, and only then, do we apply the measurement to our "unknowns". In addition, we often include positive and negative controls in each "run" or set of unknown samples to provide evidence that we have executed the method correctly each time.

In observational epidemiological studies, our instruments are statistical analyses using specific observational datasets and our analytes are specific outcomes (Table 1). Unfortunately, we rarely have the equivalent of a "reference standard" in observational studies. One possible gold standard is to compare the results from observational studies to evidence from RCTs. However, there are several limitations to RCTs as the gold standard for observational studies: there is often only a limited number of RCTs of any given topic, and their external validity may be limited [25, 26].

The steps in observational studies are not well defined. We have come to appreciate that there are a large number of decisions embedded in the application of each statistical method so, while reproducibility may be a cherished attribute of all scientific research, the sheer number of different choices makes this a difficult goal to achieve [27]. It is far too easy to be misled by spurious results or to have actually identified the outcome as a result of some misstep in our methods.

Epidemiological methods have evolved considerably from when John Snow first employed quantitative approaches to identifying the source of London's cholera outbreak. Researchers have focused considerable energy on developing methods to account for variance and have developed theories that guide their approach to accounting for bias when analyzing observational data. Using simulated data, method developers have demonstrated that, under the assumption that between/within person confounding works the way they modeled it, the method seems to adjust for it at least to a degree. The field has formulated theories that explain what epidemiological methods are most applicable (most likely to provide a good estimate of relative risk by accounting for both systematic and random error) in different situations. These theories are taught in training programs and codified in textbooks and white papers and are based on thoughtful conceptual considerations of how the world should behave [28–30]. The US Food and Drug Administration (FDA) and European Medicines Agency (EMA) endorse basic quality practices in the way observational pharmacoepidemiological studies are designed, conducted, and analyzed and communicated according to the International Society for Pharmacoepidemiology's (ISPE) Guideline on Good Pharmacoepidemiology Practices (GPP). Legislation in the European Union also mandates the use of GPP for post-approval safety studies, and currently also references the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology.

To date, however, the epidemiological research community has not fully validated its methods using a gold standard and do not include a collection of positive and negative controls in each analysis to provide evidence that we have carried out all the steps properly. Despite strong agreement that analyses should be pre-specified and efforts to require prospective reporting of analyses, retrospective observational studies analyses in particular are rarely recorded in clinicaltrials.gov with only 3,223 (July 14, 2013) out of 27,532 observational studies and only 122 (August 15, 2013) in the ENCEPP repository. Notable examples of observational analyses that are not included in clinicaltrials.gov are the FDA's mini-sentinel project studies.

In addition, the amount of detail in the description of how the analyses are implemented is quite variable. Even in the case that we thoughtfully construct an analysis in advance we depend on a degree of trial and error in actually conducting the study. We inspect the statistical distributions of the data and identify "outliers" (data values that do not seem reasonable for one reason or another) and which we remove from the analysis. We realize in looking at the dataset that there may be confounding by indication and we modify the analysis "appropriately" finding that an effect now emerges. Or, inspecting marginal totals, we notice that the effect is marked in the elderly and decide to perform a sub-group analysis of those patients. We test a large number of hypotheses and focus on the "interesting" results. We explore a number of models choosing the one that fits best. We seem to have an inherently strong capability to rationalize observations which may not serve our scientific endeavors well [31]. Epidemiologists are driven to untangle the strands in detail accounting for each

variable one by one; they attempt to accomplish what most RCTs avoid. Ideally, this would not be the case but in reality it is common—post-hoc explanations and rationalizations abound in pharmacoepidemiological studies [32]. Anyone who has attempted to replicate their methods, particularly in a close collaboration with another research group, knows how challenging recognizing much less recording these many "small" choices can be and how controversial making them can become when looking at different data sets. "Torture the data long enough and they will confess to anything" is an often quoted maxim and Young and Karr assert that "Any claim coming from an observational study is most likely to be wrong" [2]. In addition, we often apply these methods to datasets that are selected on the basis of convenience (primarily accessibility) rather than on any rational basis, even though evidence continues to emerge that the result in highly dependent on the dataset chosen in addition to the method applied [33].

Kuhn [34] asserts that when "… an existing paradigm has ceased to function adequately in the exploration of an aspect of nature to which that paradigm itself had previously led the way" the need for a paradigm shift "could be discovered only through something's first going wrong with normal research." During what Kuhn termed a "period of normal science", the failure of a result to conform to the paradigm isn't viewed as refuting the paradigm but rather attributed to an error by way of Popper's falsifiability criterion. As more inconsistent results accumulate, the field reaches a crisis and a new paradigm is established—a stage Kuhn referred to as "revolutionary science". In our view, the difficulties in replicating observational epidemiological studies has risen to the level of "something going wrong" and requires a new paradigm.

Regulatory agencies have recommended cautious and thoughtful analysis. Researchers in the field, including Suissa, suggested that qualified and experienced individuals should be designing and analyzing observational studies of drug safety [35]. We believe that relying on expertise is a weak strategy for improving our ability to replicate studies. Given the poor track record of observational studies as detailed above, we have little faith that an expert can a priori fully understand and appreciate the complexities inherent in observational data and specify an analysis that is guaranteed to produce unbiased and reliable estimates, without empirically testing afterwards whether this goal was obtained.

3 Evidence Based Epidemiology

We believe that the path forward lies not in more thoughtful and careful analysis design, but rather in augmenting current practice by applying a rigorous systematic approach in which we evaluate and improve the performance of analyses and the datasets as if they, together, constitute a process or instrument for making a measurement. Only by systematically measuring and comparing the performance of *well characterized* processes can we hope to improve our ability to measure the strength of association between drug exposure and outcome, and to distinguish between positive and negative effects. If the process is thoughtful and careful, but ad hoc, it is impossible to compare its application across different problems and therefore cannot be improved. As Popper has pointed out, a methodological approach that is not empirically based is pseudo-science (for example, assuming that the ad hoc process must work because it was carried out carefully and thoughtfully instead of empirically validating the process).

In order for a statement to be ranked as scientific, it must be possible to make an observation that conflicts with the statement. It is not possible to falsify (to make an observation that conflicts with an approach) if the approach is not empirically based since an additional "customization or modification" can always be applied that "resolves" the conflict [35]. As an illustration, if the principles and generic process by which a comparison drug should be selected cannot be clearly and explicitly stated, then any conflicting observation that might falsify the statement can be "explained" by post-hoc modification of the process for selecting a comparison drug. Prasad and Jena recently proposed a step in this direction based on routinely choosing pre-specified falsification endpoints which serves as negative controls in observational studies as has sometimes been done on an ad-hoc basis in the past [36]. Because not every negative control will have confounding and bias that is similar to the association of interest, we recommend using a large number of negative controls. These controls can then be viewed as a sample of the confounding and bias that exists in the database, and can even be used to estimate an empirical null distribution and calibrated p value [37].

3.1 Previous Research Within OMOP

In order to investigate the hypothesis that "rigorously applied, systematic processes for assessing drug–outcome associations can be constructed and compared", the Observational Medical Outcomes Partnership (OMOP) undertook an evaluation of 2,067 analyses adapted for longitudinal data to evaluate drug safety in 10 observational data sources to measure the ability of the each of the methods to discriminate between 9 positive controls and 44 negative controls [38]. In these experiments, we found that no specific analysis demonstrated superior performance

and that at traditional levels of statistical significance (RR >1, $p < 0.05$), all methods have a false positive rate >18 %, with positive predictive value <38 %. We concluded that systematic processes for risk identification can be carried out and provide useful information to supplement an overall safety assessment but that assessment of analysis performance suggests a substantial chance of identifying false positive associations. Perhaps most importantly, these findings suggested a path forward through which assessments of drug–outcome associations can be compared and improved based on empiric findings.

3.2 Current Experiments

We employed a variety of approaches to further explore the appropriate use of observational data for evaluating and identifying drug safety issues. First, in order to evaluate current practice amongst the research community, we performed a survey of the variation in analysis choices made by epidemiologists, the Epidemiology Design Decision Inventory and Evaluation (EDDIE) Survey. This survey shows experts often do not agree on what the appropriate methodology is the research a particular drug–outcome pair, underlining the need for this research [27].

Based on the initial study, we carried out additional experiments in order to gain insight into these findings, overcome limitations including the relatively small number of test cases, to provide empiric evidence on which to base selection of optimal methods for identifying drug–outcome associations and to provide data which will support interpretation of those findings. We followed an approach analogous to the initial experiments in which we systematically applied the same analyses to a number of observational datasets similarly structured in a common data model and compared performance using a set of controls which included both positive and negative controls—positive controls being drug outcome pairs for which there is consistent evidence of a causal relationship and negative controls being drug outcome pairs for which there is essentially no evidence of a direct causal relationship. We also performed experiments in which we inserted drug–outcome associations into simulated data, allowing us to explore the accuracy of odds ratio and relative risk estimates resulting from application of the methods.

First, we expanded the methods and improved the code that implemented the methods against the common data model (CDM) created in the course of the earlier set of experiments. The CDM and its associated terminologies form a foundation for systematically applying analytical methods across multiple data sources and can support drug safety and comparative effectiveness research [39, 40]. In addition to modifications which improved performance, we further parameterized the method implementations which improved maintainability, reduces complexity, and improved implementation consistency. The OMOP team created a suite of statistical methods (http://omop.org/MethodsLibrary) that encompasses the majority of analytical methods that have been employed or proposed for detection of associations between medications and clinical outcomes. We then implemented and systematically evaluated each of the seven methods: the new user cohort method [41], the case-control method [42], self-controlled case series [43], self-controlled cohort method [44], longitudinal gamma Poisson shrinker [45], temporal pattern discovery [46], and a collection of disproportionality methods [47] used in spontaneous report data. Importantly, all methods operate directly on data represented in the CDM, eliminating the need to modify methods to operate on different databases which is also a critical aspect of reproducibility. We have validated these methods through code reviews, by comparing the results obtained using these methods with those obtained in published studies and by comparing the results obtained to known signals injected in simulated data.

All analyses start with access to the entire observational database as input, and use values assigned for the relevant analysis choices to produce an output file that contains an effect estimate (e.g. relative risk) for each drug–outcome combination. All analyses are entirely algorithmic and do not require any decisions to be made by the researcher thereby minimizing variation and explicitly defining analysis process.

Second, we utilized a selection of the observational datasets that were employed in the initial set of experiments. We chose this subset based on our observations from our first series of experiments and from work done in the Exploring and Understanding Adverse Drug Reactions (EU-ADR) project [48] that databases with insufficient

Table 1 A consistent set of concept definitions to facilitate discussion of studies of statistical methods such as those conducted by OMOP

Concept	Concept definition
Method	We use the term **method** to refer to the overall method or study design (e.g. Self-Controlled Case series, Case Control, or Longitudinal Gamma Poisson Shrinker)
Analysis choices	We employ the term **analysis choices** to indicate additional, more detailed choices needed to fully specify a method, like risk window length and number of controls
Analysis	The term **analysis** is used to refer to a fully specified analysis, meaning a combination of a method with a set of specific analysis choices

Table 2 Summary of five datasets used in the second set of experiments

Name (Abbreviation)	Description	Population	Observation time	Drugs	Conditions	Procedures	Observations
Thomson MarketScan Commercial Claims and Encounters (CCAE)	Represents privately insured population and captures administrative claims with patient-level de-identified data from inpatient and outpatient visits and pharmacy claims of multiple insurance plans	Total: 46.5 m % male: 49 % Mean age [58]: 31.4 (18.1)	Patient-years: 97.6 m 2003–2009	Records: 1,030.6 m NDC from pharmacy dispensing claims HCPCS/CPT/ICD9P procedures from inpatient/outpatient medical claims	Records: 1,257.5 m ICD9 from inpatient/ outpatient medical claims	Records : 1,979.1 m HCPCS/CPT/ICD9 P procedures from inpatient/outpatient medical claims	Not available
Thomson MarketScan Multi-State Medicaid (MDCD)	Contains administrative claims data for Medicaid enrollees from multiple states, including inpatient, outpatient, and pharmacy services	Total: 10.8 m % male: 42 % Mean age [58]: 21.3 (21.5)	Patient-years: 20.7 m 2002–2007	Records: 360.2 m NDC from pharmacy dispensing claims HCPCS/CPT/ICD9P procedures from inpatient/outpatient medical claims	Records: 552.8 m ICD9 from inpatient/ outpatient medical claims	Records: 557.7 m HCPCS/CPT/ICD9 P procedures from inpatient/outpatient medical claims	Not available
Thomson MarketScan Medicare Supplemental Beneficiaries (MDCR)	Captures administrative claims for retirees with Medicare supplemental insurance paid by employers, including services provided under Medicare-covered payment, employer-paid portion, and any out-of-pocket expenses	Total: 4.6 m % male: 44 % Mean age [58]: 73.5 (8.0)	Patient-years: 13.4 m 2003–2009	Records: 400.9 m NDC from pharmacy dispensing claims HCPCS/CPT/ICD9P procedures from inpatient/outpatient medical claims	Records: 404.9 m ICD9 from inpatient/ outpatient medical claims	Records: 478.3 m HCPCS/CPT/ICD9 P procedures from inpatient/outpatient medical claims	Not available
Thomson MarketScan Lab Supplemental (MSLR)	Represents privately insured population that has at least one recorded laboratory value, with administrative claims from inpatient, outpatient, and pharmacy services supplemented by laboratory results	Total : 1.2 m % male: 35 % Mean Age [58]: 37.6 (17.7)	Patient-years: 2.2 m 2003–2007	Records: 37.6 m NDC from pharmacy dispensing claims HCPCS/CPT/ICD9P procedures from inpatient/outpatient medical claims	Records: 49.5 m ICD9 from inpatient/ outpatient medical claims	Records: 68.5 m HCPCS/CPT/ICD9 P procedures from inpatient/outpatient medical claims	Records: 41.8 m LOINC from outpatient laboratory services

Table 2 continued

Name (Abbreviation)	Description	Population	Observation time	Drugs	Conditions	Procedures	Observations
GE Centricity (GE)	Derived from data pooled by providers using GE Centricity Office (an ambulatory electronic health record) into a data warehouse in a HIPAA-compliant manner	Total : 11.2 m % male: 42 % Mean age [58]: 39.6 (22.0)	Patient-years: 22.4 m 1996–2008	Records: 182.6 m GPI from medication history and prescriptions written	Records: 66.1 m ICD9 from problem list	Records: 110.6 m CPT from procedure list	Records: 1,121.1 m LOINC for laboratory values, SNOMED for chief complaints, signs and symptoms

HIPAA Health Insurance Portability and Accountability Act of 1996, *LOINC* Logical Observation Identifiers Names and Codes, *NDC* National Drug Code, *GPI* Generic Product Identifier, *HCPCS* Healthcare Common Procedure Coding System, *CPT* Current Procedural Terminology, *ICD* International Classification of Diseases, *SNOMED* Systematized Nomenclature Of Medicine Clinical Terms

sample size for most drug–outcome pairs impair interpretation. This restriction severely limited the number of datasets that we could include in our analysis. We characterized all of these datasets (Table 2) in detail as part of the first set of experiments and the results are available on the OMOP web site. We describe a data quality framework that we used to monitor the integrity of data throughout the OMOP [49].

Third, we develop and tested multiple alternative definitions for four outcomes of interest to allow us to study the accuracy of the outcome definition and the impact on the performance of analytical methods [50]. In order to better understand the behavior of analytic methods, we examined how performance changes when different definitions of the outcomes are used [51]. We also investigated how performance depends on sample size, and for which drug–outcome pairs the sample size is sufficient in the databases used in our studies [52].

Fourth, we created a relatively large series of positive and negative controls for four health outcomes of interest (HOIs): acute kidney injury, acute liver injury, acute myocardial infarction, and upper gastrointestinal (GI) bleeding (definitions are described in [50] and are available at http://omop.org/HOI). These HOIs represent four of the most significant drug safety outcomes considered for a risk identification system [53]. For each of these outcomes, drugs were classified as positive or negative controls based on the criteria in Table 3 [54], which were chosen to, to the extent possible, include only drug–outcome pairs in which we could be highly confident were or were not associated.

Table 4 summarizes the number of test cases studied in the experiments. In our main results, we eliminated drug–outcome pairs for which the prevalence is too low in a particular database to provide adequate power.

We are not trying to determine whether the drugs and outcomes in the positive or negative controls are associated but, rather, are using these well documented associations or lack of associations as the "gold standard" against which to compare the method's performance. An ideal method would always identify an association between the drugs and outcomes in the positive control set and would never identity an association between the drugs and outcomes in the negative control set. When a method fails to identify a positive control pair as associated, the sensitivity is less than perfect and when it inappropriately identifies one of the negative control pair as being associated, the specificity is less than perfect.

We use the set of positive and negative controls to measure the performance of different methods by comparing the "gold standard" status of each drug–outcome pair with the estimated effects generated by the method when applied to the drug–outcome pair within a given observational database. We applied multiple types of

Table 3 Criteria applied for selection of positive and negative controls for the four HOIs

	Positive controls	Negative controls
Event listed in Boxed Warning or Warnings and Precautions section of the FDA Structured Product Label	True	False
Drug listed as 'causative agent' in Tisdale et al. [58]	True	False
Literature review identified no powered studies with refuting evidence of effect	True	False

Table 4 A summary of the positive and negative controls used in the second set of experiments including breakdowns by outcomes

	Positive controls	Negative controls	Total
Acute liver injury	81	37	118
Acute myocardial infarction	36	66	102
Acute renal failure	24	64	88
Upper gastrointestinal bleeding	24	67	91
Total	165	234	399

measurements in this evaluation, including area under the receiver operating characteristic curve (AUC), mean squared error, bias, and coverage probability (the probability that the 95 % confidence interval contains the true effect size). We explored the performance of methods across different databases, and within specific sub-groups of the test cases to determine if method behavior varied in different circumstances.

In order to test the reproducibility of our findings, the experiment was replicated in a collection of 6 European electronic health record databases in the EU-ADR network [55]. Finally, in order to measure the precision of estimates obtained from applying these methods we constructed several large simulated observational databases into which drug-HOI pairs with known levels of relative risk were inserted [56].

There are two major limitations of this approach. First, some of the positive and negative controls could, despite our rigorous approach, be misclassified. Even if some of the controls are misclassified as long as it is a modest percentage and there isn't a strong bias in the misclassification the limitation is not severe particularly when the comparisons between methods are made using the same positive and negative controls. Second, while the use of receiver operator characteristics is a well-established signal detection method that allows quantitative comparisons of method performance for distinguishing positive and negative controls it does not provide any assessment of the methods' ability to correctly assess the strength of association. We believe that, while it is an incomplete assessment of method performance, differentiating positive from negative associations is the critical first step and that assessing the strength of association can be approached with more confidence given a high level of certainty that an association exists. We explored a variety of alternative measures including other dichotomous measures including sensitivity, specificity, false positive rate, positive predictive value, and negative predictive value and continuous measures such as mean squared error, bias and coverage probability. We also used Hand's H-measure instead of AUC to compare methods, but this did not change our conclusions [57].

4 Conclusion

Unless the field adopts a consistent approach to analysis in which the method implementation are standardized, and the analysis and database combinations are characterized using positive and negative controls, we will not be able to fully evaluate the performance of the observational analysis process. Without such benchmark information about its empirical performance, we cannot assess whether we are improving the situation when we introduce methodological or data innovations. In order to demonstrate the feasibility of this approach, to develop initial methods for variable selection and to begin to characterize methods and databases we undertook a systematic analysis of all drugs either associated or certainly not associated with four outcomes. The papers in this supplement report the major findings from our effort to systematically apply a broad range of epidemiological methods employing a range of analytical options to a series of datasets commonly employed for observational epidemiological analyses along with associated explorations.

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. Overhage is an employee of Siemens. Drs. Ryan and Stang 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, and has become an employee of Janssen Research and Development since completing the work described here. Dr. Schuemie has previously received a grant from FNIH.

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. Popper K. Science: conjectures and refutations. In: Mace CA, editor. A lecture given at Peterhouse, Cambridge, in Summer 1953, as part of a course on developments and trends in contemporary British philosophy, organized by the British Council; originally published under the title 'Philosophy of Science: a Personal Report' in British Philosophy in Mid-Century, 1957.
2. Young SS, Karr A. Deming, data and observational studies. Significance. 2011;8(3):116–20.
3. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342(25):1887–92.
4. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med. 2000;342(25):1878–86.
5. Danaei G, Tavakkoli M, Hernán MA. Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. Am J Epidemiol. 2012;175(4):250–62.
6. MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technol Assess. 2000;4(34):1–154.
7. Furlan AD, Tomlinson G, Jadad AA, Bombardier C. Methodological quality and homogeneity influenced agreement between randomized trials and nonrandomized studies of the same intervention for back pain. J Clin Epidemiol. 2008;61(3):209–31.
8. Abraham NS, Byrne CJ, Young JM, Solomon MJ. Meta-analysis of well-designed nonrandomized comparative studies of surgical procedures is as good as randomized controlled trials. J Clin Epidemiol. 2010;63(3):238–45.
9. Suissa S. Randomized trials built on sand: examples from COPD, hormone therapy, and cancer. Rambam Maimonides Med J. 2012;3(3):e0014.
10. Tuma RS. Statisticians set sights on observational studies. J Natl Cancer Inst. 2007;99(9):664–5, 8.
11. Varas-Lorenzo C, Garcia-Rodriguez LA, Perez-Gutthann S, Duque-Oliart A. Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case–control study. Circulation. 2000;101(22):2572–8.
12. Grodstein F, Stampfer MJ, Manson JE, Colditz GA, Willett WC, Rosner B, et al. Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. N Engl J Med. 1996;335(7):453–61.
13. Wilson PW, Garrison RJ, Castelli WP. Postmenopausal estrogen use, cigarette smoking, and cardiovascular morbidity in women over 50. The Framingham Study. N Engl J Med. 1985;313(17):1038–43.
14. Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, et al. Estrogen plus progestin and the risk of coronary heart disease. N Engl J Med. 2003;349(6):523–34.
15. Hernán MA, Robins JM, García Rodríguez LA. Discussion on "Statistical Issues Arising in the Women's Health Initiative". Biometrics. 2005;61(4):922–30.
16. Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008;19(6):766–79.
17. Cardwell CR, Abnet CC, Cantwell MM, Murray LJ. Exposure to oral bisphosphonates and risk of esophageal cancer. JAMA J Am Med Assoc. 2010;304(6):657–63.
18. Green J, Czanner G, Reeves G, Watson J, Wise L, Beral V. Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. BMJ. 2010;341:c4444.
19. Meier CR, Schlienger RG, Kraenzlin ME, Schlegel B, Jick H. HMG-CoA reductase inhibitors and the risk of fractures. JAMA J Am Med Assoc. 2000;283(24):3205–10.
20. van Staa TP, Wegman S, de Vries F, Leufkens B, Cooper C. Use of statins and risk of fractures. JAMA J Am Med Assoc. 2001;285(14):1850–5.
21. de Vries F, de Vries C, Cooper C, Leufkens B, van Staa TP. Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the General Practice Research Database. Int J Epidemiol. 2006;35(5):1301–8.
22. Ioannidis JP, Tzoulaki I. Minimal and null predictive effects for the most popular blood biomarkers of cardiovascular disease. Circ Res. 2012;110(5):658–62.
23. McDonald CJ. The evolution of Intel's copy exactly! Technology transfer method. Intel Technol J. 1998;Q4:1–6.
24. Terwiesch C, Xu Y. The copy exactly ramp-up strategy: trading-off learning with process change. August 4, 2003, cited 2012 December 24. http://qbox.wharton.upenn.edu/documents/opim/research/P6.pdf.
25. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82–93.
26. Rothwell PM. Factors that can affect the external validity of randomised controlled trials. PLoS Clin Trials. 2006;1(1):e9.
27. Stang PE, Ryan PB, Overhage JM, Schuemie MJ, Hartzema AG, Welebob E. Variation in choice of study design: findings from the epidemiology design decision inventory and evaluation (EDDIE) survey. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0103-1.
28. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Guide on methodological standards in pharmacoepidemiology (revision 1). EMA/95098/2010. Cited 2013 January 23. http://www.encepp.eu/standards_and_guidances/documents/ENCePPGuideofMethStandardsinPE.pdf.
29. Gagne JJ, Fireman B, Ryan PB, Maclure M, Gerhard T, Toh S, et al. Design considerations in an active medical product safety monitoring system. Pharmacoepidemiol Drug Saf. 2012;21(Suppl 1):32–40.
30. Gagne JJ, Nelson JC, Fireman B, Seeger JD, Toh D, Gerhard T, et al. Taxonomy for monitoring methods within a medical

product safety surveillance system: year two report of the mini-sentinel taxonomy project workgroup (workgroup) 2012, cited 2012 October 29. http://www.mini-sentinel.org/work_products/Statistical_Methods/Mini-Sentinel_Methods_Taxonomy-Year-2-Report.pdf.
31. Taleb NN. The Black Swan: the impact of the highly improbable. New York: Random House; 2010.
32. Avorn J. In defense of pharmacoepidemiology—embracing the yin and yang of drug research. N Engl J Med. 2007;357(22):2219–21.
33. Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol. 2013;178(4):645–51.
34. Kuhn TS. The structure of scientific revolutions. 3rd ed. Chicago: University of Chicago Press; 1996.
35. Suissa S. Time-related biases in pharmacoepidemiology. In: International Society of Pharmacoepidemiology mid-year meeting, Miami Beach, Florida, 2012.
36. Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? JAMA J Am Med Assoc. 2013;309(3):241–2.
37. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med. 2013. doi:10.1002/sim.5925.
38. 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.
39. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inf Assoc JAMIA. 2012;19(1):54–60.
40. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inf. 2012;45(4):689–96.
41. 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 (in this supplement issue). doi:10.1007/s40264-013-0099-6.
42. Madigan D, Schuemie MJ, Ryan PB. Empirical performance of the case–control method: lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0105-z.
43. 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 (in this supplement issue). doi:10.1007/s40264-013-0100-4.
44. 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 (in this supplement issue). doi:10.1007/s40264-013-0101-3.
45. Schuemie MJ, Madigan D, Ryan PB. Empirical performance of LGPS and LEOPARD: lessons for developing a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0107-x.
46. Norén GN, NG, 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 (in this supplement issue). doi:10.1007/s40264-013-0095-x.
47. DuMouchel B, Ryan PB, Schuemie MJ, Madigan D. Evaluation of disproportionality safety signaling applied to healthcare databases. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0106-y.
48. 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 Saf. 2012;21(6):611–21.
49. Hartzema AG, Reich C, Ryan PB, Stang PE, Madigna D, Welebob E, et al. Managing data quality for a drug safety surveillance system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0098-7.
50. 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 (in this supplement issue). doi:10.1007/s40264-013-0104-0.
51. Reich C, Ryan PB, Schuemie MJ. Alternative outcome definitions and their effect on the performance of methods for observational outcome studies. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0111-1.
52. 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 (in this supplement issue). doi:10.1007/s40264-013-0112-0.
53. Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salame G, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiol Drug Saf. 2009;18(12):1176–84.
54. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0097-8.
55. 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 (in this supplement issue). doi:10.1007/s40264-013-0109-8.
56. Ryan PB, Schuemie MJ. Evaluating performance of risk identification methods through a large-scale simulation of observational data. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0110-2.
57. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn. 2009;77(1):103–23.
58. Tisdale J, Miller D. Drug-induced diseases: prevention, detection, and management. 2nd ed. USA: American Society of Health-System Pharmacists; 2010.