Big Data, Big Results: Knowledge Discovery in Output from Large-Scale Analytics

Tyler H. McCormick[1]*, Rebecca Ferrell[1], Alan F. Karr[2] and Patrick B. Ryan[3]

[1]*Department of Statistics, University of Washington, Seattle, WA 98195, USA*

[2]*National Institute of Statistical Sciences, Research Triangle Park, NC 27709, USA*

[3]*Janssen Research and Development & OMOP, Titusville, NJ 08560, USA*

Received 1 July 2013; revised 21 May 2014; accepted 24 May 2014
DOI:10.1002/sam.11237
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Observational healthcare data, such as electronic health records and administrative claims databases, provide longitudinal clinical information at the individual level. These data cover tens of millions of patients and present unprecedented opportunities to address such issues as post-market safety of medical products. Analyzing patient-level databases yields population-level inferences, or 'results', such as the strength of association between medical product exposure and subsequent outcomes, often with thousands of drugs and outcomes. In this article, by contrast, we study 'big results', which are the product of applying thousands of alternative analysis strategies to five large patient databases. These results were produced by the Observational Medical Outcomes Partnership. All together, there are more than 6 million results, comprising risk assessments for 399 medical product–outcome pairs analyzed across five observational databases using seven statistical methods, each of which has between a few dozen and a few hundred variants representing parameters or 'tuning variables'. We focus on the value of knowledge discovery methods and the challenges in extracting clinically relevant knowledge from big results. We believe our analyses are both scientifically and methodologically valuable as they reveal information about how methods/algorithms perform under various circumstances, as well as provide a basis for comparison of these methods. © 2014 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 7: 404–412, 2014

Keywords: big results; clustering; partitioning; results mining; drug safety; observational medical data

1. INTRODUCTION

Administrative healthcare data systems now provide a source of rich longitudinal observational data from tens of millions of patients, covering important clinical elements of health service utilization such as test results, disease diagnoses, and drug utilization. From these data have come tremendous advances in high-throughput computing and machine-learning algorithms that now allow the systematic exploration of large health datasets for novel patterns.

In the interest of public health at a population-level, one such set of patterns is the temporal relationship of the incidence of adverse health outcomes following drug exposure. The research community, with contributions from epidemiology, statistics, computer science, informatics, and machine learning, has devised hundreds of alternative algorithms that can mine observational data to estimate the strength of association between exposure and outcome. In any given observational database of health information, there are tens of thousands of different product exposures to investigate and thousands of potential outcomes resulting from those exposures. Assuming a network of 10 disparate databases, this means the result set from mining for potential associations can exceed: $10 \times 100 \times 10\,000 \times 1000 = 10\,000\,000\,000$ observations.

Each such database can be analyzed in multiple ways, and most analysis methods have multiple variants. For instance, how soon following administration of a drug must an adverse outcome occur in order to be considered associated with the drug? Such multiple analyses of 'big data' sources themselves generate big data, which we term 'big results'. Methods have not been established for evaluating

* *Correspondence to:* Tyler H. McCormick (tylermc@uw.edu)

big results, and there are no agreed standards for how appropriately to use big results. For example, which algorithms are most appropriate, under what circumstances? How well do these algorithms work (predictive accuracy)? How consistent are results across different databases or different drug–outcome scenarios? How can analysis results be successfully combined to produce more meaningful composite summaries?

The issues we address in this article arise as immediate consequences of novel mechanisms to collect massive amounts of observational data. The data we use, from the Observed Medical Outcomes Partnership (OMOP), demonstrate the potential value of massive observational data in healthcare research. Similar data are increasingly present in other domains, such as finance, marketing, social media, and the social sciences. One challenge in analyzing big results arises from a mismatch between standard statistical analysis paradigms and the nature of massive observational data. In the healthcare context, for instance, standard results for hypothesis testing assume that an association of interest, statistical analysis tool, and desired effect size are chosen *a priori*. There has been much work in the statistics literature on developing adjustments and robust tools to account for testing multiple associations or outcomes of interest (e.g., the classical multiple testing problem in a genome-wide association study), but these methods typically assume that the analysis framework is set beforehand in a reasonably controlled sample. In observational contexts, especially as new methods are devised, it is difficult to develop the necessary intuitions to make such decisions in advance. Further, multiplicity adjustment essentially assumes the problem is randomness such that Type I error control can help minimize risk of spurious findings. In observational data, we cannot assume unbiased estimators; we know that the data are replete with potential sources of bias. What we do not know is how well-calibrated methods are in their adjustment of bias, or how well they work across a range of different problem spaces where bias may vary (e.g., different drugs and different outcomes, or different databases).

In this article, we report initial explorations of the big results problem. We also describe a specific application in medical product surveillance and show some exploratory analysis methods applied to big results from the OMOP experiment. The primary contribution of this article is in formulating the big results problem, providing context by drawing parallels with related problems in statistics and machine learning, and demonstrating the value of data mining techniques in formulating clinically relevant knowledge in this domain. In the remainder of this section, we introduce the OMOP result set and discuss three questions that we seek to answer. In Section 2, we demonstrate several statistical approaches to address these questions. Finally, Section 3 discusses the big results problem more generally, providing suggestions for the data mining community.

1.1. Big Results: The OMOP Result Set

One fundamental purpose of OMOP is to catalyze development of research methods that inform the appropriate use of observational healthcare data in studying the effects of medical procedures and products. To achieve this goal, OMOP designed a series of empirical experiments to test alternative statistical methods across a network of observational health databases, and to measure their performance across an array of different medical product exposures and health outcomes of interest. The data from this experiment (the result set) and all research materials, including source code for the methods and documentation, are publicly available and can be downloaded from http://omop.org.

For the OMOP experiment, four health outcomes were considered: acute myocardial infarction, acute liver injury, acute renal failure, and gastrointestinal bleeding. For each outcome, particular drugs were selected, each of whose association or nonassociation of the drug with the outcome was known beforehand—there is a 'ground truth'. In all, 399 combinations of drugs and outcomes (drug–outcome pairs) were analyzed by the OMOP. These pairs were classified as 'positive controls' when there was prior evidence from external sources (e.g., clinical trials, spontaneous reports, and literature) that the drug increased the risk of the outcome (e.g., naproxen and gastrointestinal bleeding) and as 'negative controls' when no evidence was available to suggest a positive association (for instance, ketoconazole and acute liver injury). Given a lack of consensus in the literature about how to identify these particular outcomes in observational databases, which may be based on diagnosis codes, therapeutic procedures, laboratory results, or other information contained in health data, multiple definitions of the outcomes were examined for each drug–outcome pair.

Seven well-known statistical methods were used to evaluate the association between drug–outcome pairs. These seven methods, listed in Table 1, are all commonly used in the context of drug safety. The methods make use of observational patient-level data in different ways. The Self-Controlled Case Series method [1], for instance, uses data from a given patient when the patient is not taking a drug to compute a baseline patient-specific hazard rate. The Longitudinal Gamma Poisson Shrinker [2], by contrast, uses the exposure profiles of patients with given characteristics across the entire observation window to compute a baseline risk.

Within each of the seven methods, the OMOP team designed multiple variants corresponding to plausible

Table 1. Abbreviations for the seven methods included in the OMOP result set. Additional details about methods are available at http://omop.org/sites/default/files/Methods%20in%20OMOP%202011_2012%20Research.pdf. An expanded list of publications related to these methods is available at http://omop.org/sites/default/files/OMOP%202012%20symposium%20citations_FINAL.pdf.

Method abbreviation	Method name	Publication
CC	Case Control	[6]
CM	Cohort Method	[6]
DP	Disproportionality Analysis	[6]
ICTPD	IC Temporal Pattern Discovery	[7]
LGPS	Longitudinal Gamma Poisson Shrinker	[2]
OS	Observational Screening	[8]
SCCS	Self-Controlled Case Series	[1]

potential tuning settings that a researcher might consider in evaluating drug–outcome associations. For some methods, such as the Self-Controlled Case Series, different variants result from, for instance, using different values for hyperparameters on prior distributions. For other methods, a variant may refer to adjusting the amount of time between drug exposures required to count an observation as a new exposure (commonly known as a 'wash-out period'). The number of variants examined within a method ranges from roughly 20 variants to about 150. A higher number of variants cannot, however, be assumed to lead to more variability in risk estimates. The variants instead correspond to a subset of the universe of possible parameterizations that might be employed for a method in real-world drug safety research.

OMOP relied on five large observational health databases, each containing millions of patients of various ages, genders, insurance programs, and other demographic attributes within the United States (e.g., the MarketScan® Commercial Claims and Encounters database). Each of the 399 drug–outcome pairs was evaluated across all OMOP databases using all variants of the seven statistical methods and all applicable definitions of the outcomes. Standardized programs implementing these analyses in each database were developed to estimate the strength of association between drug exposure and an outcome, outputting an effect measure—we focus on log relative risk—and associated standard error. These standard outputs were assembled for all 399 drug–outcome pairs and used as estimates to be compared against reference standards for descriptive characterization of the findings. The entire big results dataset consists of 6 214 822 records indexed by patient database, drug–outcome pair (with associated ground truth), and analysis method (with parameters).

Additionally, for each drug–outcome pair and database, a minimum detectable relative risk was calculated to serve as a measure of data availability analogous to power and sample size considerations, as appropriate in the context of retrospective data. The minimum detectable relative risk was calculated by estimating the prevalence of the drug exposure and outcome occurrence in each age decile-by-gender strata, calculating the number of expected events assuming independence, and then aggregating across strata. The composite expected events is then used to estimate the minimum detectable relative risk, assuming a standard cohort design, with Type I error $= 0.05$ and Type II error $= 0.80$. The specific approximation applied is discussed in ref. [3]. In analyzing data from the OMOP experiment, we use a minimum detectable relative risk of ≤ 1.25 to exclude cases without sufficient sample to be reliably studied in a given database. Thus, our big OMOP result set consists of (log) relative risk measurements, confidence intervals, and minimum detectable relative risks computed across a number of dimensions to identify associations between drugs and outcomes for which the ground truth of increased risk or no increased risk is already known.

The OMOP experiment measured operating characteristics of methods across different databases and outcomes through standard metrics. OMOP evaluated predictive accuracy of discriminating positive controls and negative controls through the area under ROC curve (AUC). The results were also used to characterize the distribution of estimates among negative controls, assuming true relative risk of 1, i.e. the empirical null distribution. OMOP also measured coverage probability in both real data and simulation, to estimate the proportion of scenarios in which a method's confidence interval contains the true effect size. In an ideal setting, the true effect sizes of all drug–outcome pairs would be known, in which case the performance of method estimates could be measured through mean squared error. Unfortunately, the true effect sizes are unknown in the real-world population, and cannot be readily approximated (except for negative controls). Instead, OMOP measured mean squared error only through simulation studies, which are outside the scope of this discussion. Further description of OMOP is available elsewhere [4]. Results from initial OMOP experiments have been published previously [5].

In total, the OMOP result set contains approximately 6 million log relative risk scores and confidence intervals, representing a collection of results on an unprecedented scale for observational data research. We approached the OMOP result set as an opportunity for knowledge discovery, in part because the abstractions needed for detailed statistical analysis remain unclear. Our primary data mining goal with the OMOP result set is to identify methods that accurately reveal dangerous drug–outcome associations without being misled by spurious associations. The most common approaches to drug safety focus on characterizing the relationship between a particular drug and specific outcome and are often initiated based on some prior hypothesis about the nature of the relationship. Such

data lead to a specific statistical test to make inferences about a (typically univariate) quantity of interest. The OMOP electronic health record databases, in contrast, contain observational data collected from multiple sources and present the opportunity to investigate multiple drugs and outcomes at the same time. The question from the OMOP result set is not whether or not a drug–outcome association exists, but rather if the choice of statistical method, data source, or tuning parameter is likely to impact the ability to detect an association.

The OMOP result set was originally developed to characterize method performance. However, we recognized that the same result set could further our understanding of method behaviors. To make this problem more concrete, we discuss the following three questions arising from our interest in variation and concordance among the observations in the result set:

1. How much do choices about product, outcome, database, and method contribute to the overall variability in the OMOP result set?
2. Which methods are generally most well-calibrated? How much variability in calibration exists across and within methods?
3. How often do methods suggest similar a clinical decision, accounting for statistical uncertainty?

We believe these three questions provide knowledge that can inform clinical decision making based on the OMOP result set. The first question provides a baseline exploration of the sources of variability in the OMOP result set. The second question investigates calibration, which in this context we take to mean a method's ability to appropriately reject (or not) false (or not) null hypotheses. The third question facilitates comparing methods while accounting for statistical uncertainty.

2. KNOWLEDGE DISCOVERY IN BIG RESULTS

In this section, we use the three questions posed in the previous section to motivate three examples of how to analyze the OMOP big results dataset.

2.1. Sources of Variation

The observations in the result set are different ways of estimating the same relative risk for a given drug–outcome. The design of the OMOP experiment provides several dimensions along which relative risk point estimates vary (and indeed, which collectively explain all variation in relative risk estimates, as there is no source of measurement error). An obvious question of interest is simply which dimensions of the results set explain the variability in relative risk estimates the most.

A simple method to better understand sources of variability in the result set is based on decompositions of sums of squares. Within each combination of drug–outcome definition, and method, we examine the sources of variation in the log relative risks by calculating the proportion of sum of squares accounted for by (i) database analyzed and (ii) specific method variant. For each of these components, *marginal* (sometimes known as 'type III') sums of squares were computed within a combination ($SS_{DB|method}$ and $SS_{method|DB}$) (to remove the effect of the sequence of fitting variables) and expressed as a proportion of the total sum of squares—essentially, the increase in R^2 obtained by adding one of these variables to a linear model already including the other.

We report the empirical densities of each of these proportions across all such combinations of drug–outcome definition, and method in the result set in Fig. 1. Figure 1 shows that the proportion of variation between relative risk estimates within these groups associated with the database tends to be low across all four outcomes and for most methods except for LGPS and DP. That is, under this simple linear framework, we typically gain relatively little ability to predict relative risk in the result set by knowing which database was analyzed once we know the tuning parameters used, with the exception of the LGPS and DP methods. Correspondingly, the proportion of variation associated with the specific implementation of each method after accounting for database analyzed has a flatter distribution and takes on values closer to 1 more frequently. This increase in R^2 from knowing the method variants tended to be small for LGPS and larger for the other types [particularly Self-Controlled Case Series (SCCS)]. From this we conclude that we typically gain more ability to predict relative risk in the result set by knowing which tuning parameters were used than which database was analyzed. We also note that across all four OMOP outcomes, the distributions of sources of variability are similar: the variation in relative risk estimates accounted for by database or by method variant appears independent of the particular outcome analyzed.

2.2. Exploring Calibration through Tree Models

In this section, we show how exploratory analyses using machine learning begin to illuminate the differences among the statistical methods in the OMOP method result set. We stress that not only the scale, but also the very nature of the data is unusual: the unit of analysis is 'statistical method A with parameter settings B applied to (drug–outcome) pair C in database D'. Without being able to rely on

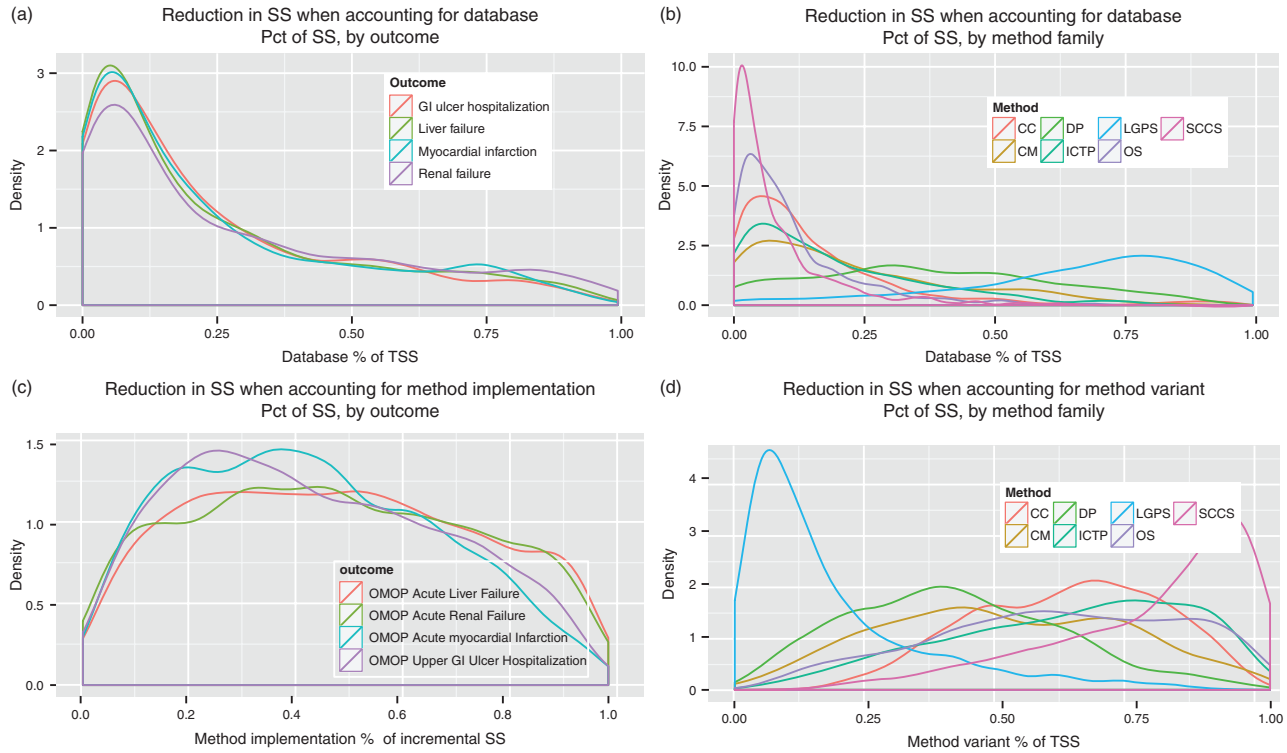

Fig. 1 Distribution of proportion of variation explained by each of primary dimensions of variation of interest. Reduction in percentage of sum of squares (SS) when accounting for (a) database, by outcome, (b) database, by method family, (c) method variant, by outcome, and (d) method variant, by method family. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the formal statistical frameworks that would follow if we had a clearer articulation of the broader sample space and sampling scheme, a machine-learning approach is appealing or even mandatory.

To perform the analyses, we restricted the result set to negative controls (ground truth = 0), meaning that the drug is believed not to be associated with the outcome. This decision was made so that bias away from a true log relative risk of 0 would be apparent, unlike in the case of the positive controls in which bias away from a positive but unknown log relative risk differing for each drug–outcome pair would be less readily detectable. We further restricted the data to observations in which the minimal detectable relative risk is greater than 1.5 to eliminate instances with low statistical power resulting from small sample size. A total of 1 953 660 cases remained. Within this filtered set of results, there are 219 (drug–outcome) pairs. There are seven classes of statistical analysis methods, each of which has multiple settings for several parameters. The methods are labeled as shown in Table 1.

We first compared the behavior of log relative risk across the methods, separately for the 219 (drug–outcome) pairs. Figure 2 shows one of the 219 sets of box plots and kernel density estimates. Because the ground truth is zero for all cases, log relative risk should be near or below zero, but often it is not. The spread within each method results from variation over the five databases, which exists but is minor, as well as from variation of the method parameters, whose effect is decidedly not minor. For this particular case, only one method (SCCS) is properly centered at zero, most methods are biased upward or downward from zero, and for some—notably case control (CC)—the variation resulting from analytical choices is substantial. The statistical implications of this variation are sobering: by setting the values of parameters, which are often not reported in journal articles, the log relative risk can take essentially any value from −1 to +4.

To understand whether the methods vary systematically, we applied partitioning [9] to each of the 219 (drug–outcome) pairs. In each case, we made only the most significant split of the even methods into a ‘low relative risk’ subset and its complementary ‘high relative risk’ subset. Figure 3 shows the results for the same (drug–outcome) pair as in Fig. 2. In this case, only the IC Temporal Pattern Discovery (ICTPD) method is low, and the mean log relative risks differ between the high and low sets by approximately 0.75. The remaining methods produce large numbers of false positives.

Finally, we conducted exploratory analyses of this 219-point summary data set. Table 2 shows that the methods

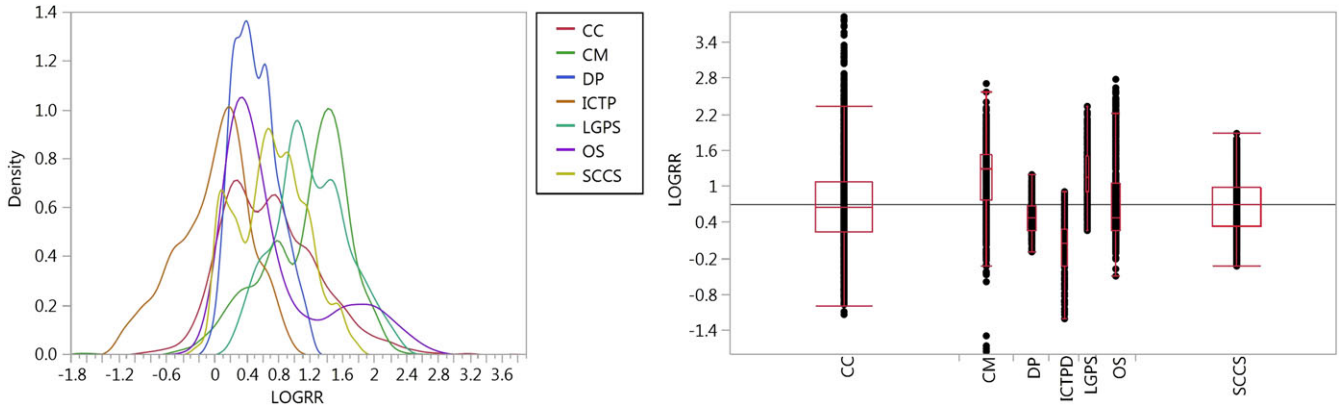

Fig. 2 Comparison of the distribution of log relative risk for seven analysis methods for one (drug–outcome) pair. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

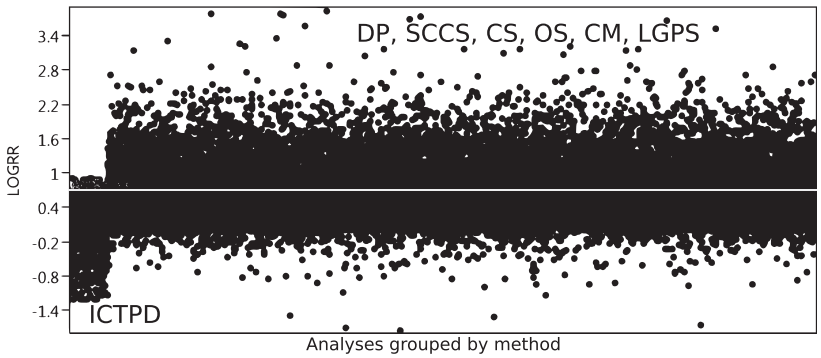

Fig. 3 Partition of the seven analysis methods into 'Low' log relative risk and 'High' log relative risk for one (drug–outcome) pair.

differ dramatically. For instance, methods CM, DP, and ICTPD are consistently in the low group, while methods LPGS and OS are generally in the high group. Only CC and SCCS are in the low and high groups with approximately equal frequencies.

Even more striking is that for among the $2^7 - 2 = 126$ possible partitions of the 219 cases into low and high subsets, 104 cases (nearly one-half of the data!) are one of eight possibilities: Low = {CM, DP}, Low = {CC, CM, DP, ICTPD, SCCS}, Low = {CM, ICTPD}, Low = {CM, DP, ICTPD}, Low = {CM, DP, ICTPD, SCCS}, Low = {CM}, Low = {CC, CM, DP} and Low = {CC, CM, DP, SCCS}. Clearly, there are both systematic similarities and systematic differences among the methods. There are also to-be-resolved subtleties: even though in 16% of the cases, method ICTPD is the unique low method, in Fig. 3, its mean log relative risk is essentially zero. The implications for statistical issues such as calibration are now being studied.

2.3. Clustering with Uncertainty

Clustering the OMOP analyses—that is, the variants of each method included—is another attempt to make sense of the big results dataset. The goal is to understand which analyses tend to produce similar results in that they are insensitive to the particular database, drug, and outcome definition used for a given outcome.

Table 2. Numbers of times in 219 (drug–outcome) pairs that each method was in the 'Low' or 'High' group.

Method	Cases in low	Cases in high
CC	96	123
CM	146	73
DP	155	64
ICTPD	93	113
LGPS	26	193
OS	25	194
SCCS	106	113

We consider a crude measure of 'correctness' of an analysis based on the ground truth for the drug–outcome pair and the 95% confidence interval for relative risk. For negative control drug–outcome pairs (ground truth of zero), a 95% interval containing 1 is considered to have obtained a 'correct' result. For positive control pairs (ground truth of 1), a 95% interval with a lower bound relative risk above 1 is considered to have obtained a 'correct' result.

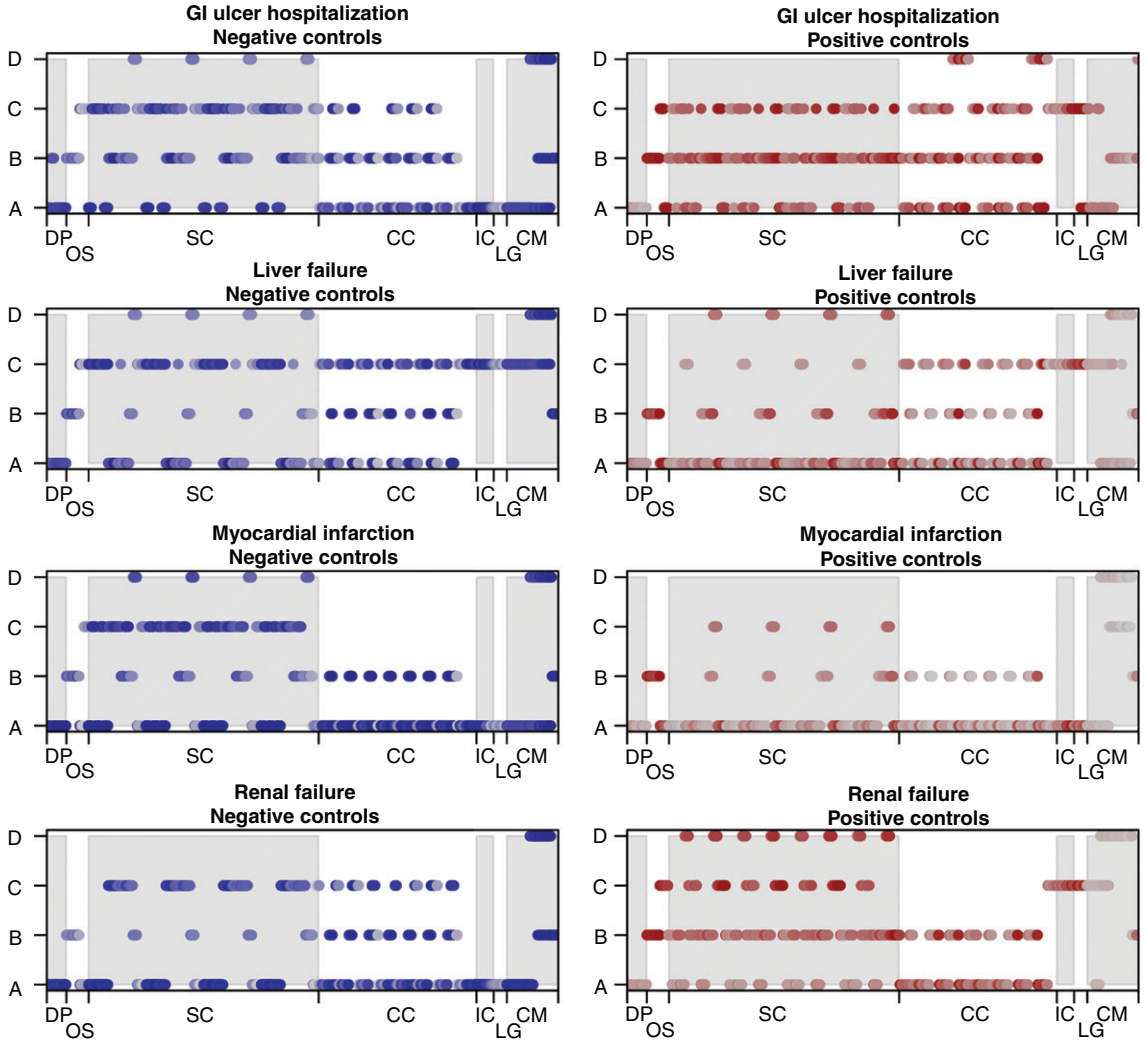

Fig. 4 Clusters of analyses within sets of outcomes and ground truths with similar performances (as defined by ground truth captured in the 95% CI). Color intensity indicates better overall performance. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Within a set of outcomes and ground truths in OMOP, a given analysis was repeated hundreds of times across multiple outcome definitions, drugs, and data sources. The similarity in performance of two analyses within a set of outcomes and ground truths can be measured by comparing the proportion of runs in which they agreed on a result, i.e. the two analyses were either both 'correct' or both 'incorrect' for a particular drug–outcome definition-database combination according to the confidence interval-based definition above.[1]

$$d(a,b) = 1 - \frac{\#\text{ common results with } a \text{ and } b \text{ both 'correct' or 'incorrect'}}{\#\text{ results common to } a \text{ and } b} \quad (1)$$

[1] Excluding analyses with minimum detectable relative risk over 1.25.

Using this rough measure of dissimilarity, we performed hierarchical clustering on the set of analyses within each of the eight outcome-ground truth combinations, partitioning the analyses into four clusters as shown in Fig. 4. Figure 4 demonstrates that variants within a family of methods (marked along the horizontal axes in the order of their analysis identifier–neighboring analyses are generally variants differing by one attribute) performed similarly (i.e., grouped into the same clusters along the vertical axes) as would be expected. However, the strength of this relationship varied among outcomes and ground truths from relatively little ability to distinguish methods for GI ulcers to clear separation in the renal failure positive controls. Also, within an outcome, analyses tend to fall into the same groups for both positive and negative controls, but the accuracy of results grouped together—depicted by the

darkness of the point colors—is approximately inversely related between the negative and positive controls. For example, the same variants of the CM method appear to be relatively strong at correctly predicting negative controls and relatively poor at predicting positive controls. In the context of mining this 'big results' data, this type of machine-learning method assists in honing in on particular groups of analyses for further inspection, such as the relatively strong performing variants of the SCCS method in clusters C and D for acute renal failure.

3. DISCUSSION

In this article, we applied several data mining approaches to reveal links between exploratory and confirmatory data analyses. We analyzed a database of about 6 million results generated using observational medical data.

In addressing our first question about sources of variability, we found that a small proportion of the total variance was explained by database type, while variants of the methods examined explained most of the total variation. This result is somewhat surprising since the demographic characteristics of patients vary substantially across databases (patients in the Medicare database tend to be older, for example). In Madigan *et al*. [10], the authors examined variability in 53 drug–outcome pairs across 10 databases within two particular method variants. Madigan *et al*. found substantial disagreement across databases in terms of direction of relative risk and statistical significance for these drug–outcome pairs, including cases in which statistical significance pointed in opposite directions. The problem of high variability across databases identified in ref. [10] combined with the even larger variability across method variants found in our investigation suggests taking extreme caution in using electronic healthcare databases to make general statements about drug safety.

Our second question explores variability within methods further and finds substantial differences in log relative risk both between methods and for different variants of a given method. This finding implies that further development in sensitivity assessment and calibration is required, especially for the methods with overall higher relative risk under zero controls. Our third question deals specifically with variability over the space of decisions that would be made using different method variants. We found that the consistency across method variants depends on the outcome of interest.

An immediate implication for infusion involves the reporting of results from drug safety studies and observational medical data more generally. Our findings underscore the importance of transparency and complete specification and reporting of analyses, as all study design choices were shown to have the potential to substantially shift effect estimates. The framework presented will also become increasingly relevant as observational medical data are used to test an increasingly wide range of associations (see ref. 11 for a recent example), with the same strategies being useful in domains outside of healthcare where big result sets are also increasingly common.

Observational healthcare data are only one example of big data that are actively being used to explore temporal patterns in longitudinal data. Clickstream analysis examines patterns of user activity for potentially novel applications such as mapping the space of scientific disciplines using logs from academic journal websites [12]. Mining text data from social media and blogs for public health trends (such as influenza outbreaks) is another area of research [13]. In all of these cases, big data offer the opportunity to analyze a nearly infinite space of relationships. However, this article raises questions about the validity of those analyses, and the degree to which evidence generated from big data exploratory analyses can be relied upon for decision making. Evaluating method validity is commonplace in statistics for specific problem spaces, but it could be argued that the dawn of big data represents the infancy of method validation for big data. The future requires generation and evaluation of big results in order to achieve validation. Big results offer the opportunity to explore not only the intended behavior of methods within a defined context, but also the complex relationships among methods and databases and across an array of problems outside of where the analyses was originally designed. In this case, big results offer the opportunity to draw inferences about method behavior in much the same way big data offer the opportunity to draw inferences about individual behavior. In this regard, analysis of big results can be considered a machine-learning problem and may utilize machine-learning techniques, but further work is required to establish best practices for defining the appropriate big results questions to ask and identifying the appropriate big results analyses to employ.

An additional lesson involves the importance of the data generation mechanism in new forms of observational data. Throughout our analyses, we were continually confronted with questions about what portion of the total universe of, e.g. drug–outcome pairs, or of analyses, is in our ground truth dataset. From a machine-learning perspective one might cast knowledge discovery in a big result set as a prediction problem as in the meta-learning literature (see, for instance, ref. 14). That is, we may build models where we predict the association (measured in log relative risk) between a drug and outcome as a function of the database and analysis method:

$$\log \mathrm{RR}_{d_i,o_j} = f(\text{database, method}) + \epsilon \qquad (2)$$

for drug d_i and outcome o_j. Though it is computationally feasible to build such models, scientifically it is less clear that the performance of a method on one drug–outcome pair will be associated with the performance on a different pair. Further, we do not know which corner of the universe of drug–outcome pairs we observe in the OMOP database. Doing out-of-sample prediction using other drugs in the database, therefore, does not yield decisive evidence of how the method will perform on a new drug–outcome pair. Analyses evaluated using the ground truth information have similar issues, because the universe of true associations is, of course, unknown and it is unclear when (or if) more ground truths will become available. Without knowledge of how (or if) representative our selection of drug–outcome pairs are of the all possible pairs, our ability to convey our results hinges on carefully thinking through the implications of various possible types of bias.

ACKNOWLEDGMENTS

This research was supported in part by NSF grant DMS–1127914 to the Statistical and Applied Mathematical Sciences Institute (SAMSI). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] H. J. Whitaker, M. N. Hocine, and C. P. Farrington, The methodology of self-controlled case series studies, Stat Methods Med Res 18(1) (2009), 7–26.

[2] M. J. Schuemie, Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD, Pharmacoepidemiol Drug Saf 20(3) (2011), 292–299.

[3] B. Armstrong, A simple estimator of minimum detectable relative risk, sample size, or power in cohort studies, Am J Epidemiol 126(2) (1987), 356–358.

[4] P. E. Stang, P. B. Ryan, J. A. Racoosin, J. M. Overhage, A.G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock, Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership, Ann Intern Med 153 (2010), 600–606.

[5] P. B. Ryan, D. Madigan, and P. E. Stang, Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership, Stat Med 31 (2012), 4401–4415.

[6] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman, Novel data-mining methodologies for adverse drug event discovery and analysis, Clin Pharmacol Ther 91(6) (2012), 1010–1021.

[7] G. N. Noren, A. Bate, J. Hopstadius, K. Star, and I. R. Edwards, Temporal pattern discovery for trends and transient effects: its application to patient records, In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Las Vegas, NV, 2008, 963–971.

[8] P. B. Ryan, G. E. Powell, E. N. Pattishall, and K. J. Beach, Performance of Screening Multiple Observational Databases for Active Drug Safety Surveillance, Providence, RI, International Society of Pharmacoepidemiology, 2009.

[9] L. Breiman, Classification and Regression Trees, Boca Raton, FL, Chapman & Hall/CRC, 1984.

[10] D. Madigan, P. B. Ryan, M. Schuemie, P. E. Stang, J. M. Overhage, A. G. Hartzema, M. A. Suchard, W. DuMouchel, J. A. Berlin, Evaluating the impact of database heterogeneity on observational study results, Am J Epidemiol 178(4) (2013), 645–651.

[11] M. Schuemie, P. Coloma, H. Straatman, R. Herings,G. Trifir, J. Matthews, D. Prieto-Merino, M. Molokhia, L. Pedersen, R. Gini, Innocenti, F., Mazzaglia, G., G. Picelli, L. Scotti, J. van der Lei, M. Sturkenboom, Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods, Med Care 50 (2012), 890–987.

[12] J. Bollen, H. Van de Sompel,A. Hagberg, L. Bettencourt, R. Chute, M. A. Rodriguez, and L. Balakireva, Clickstream data yields high-resolution maps of science, PLoS ONE 4(3) (2009), e4803.

[13] C. D. Corley, D. J. Cook, A. R. Mikler, K. P. Singh, Text and structural data mining of influenza mentions in web and social media, Int J Environ Res Public Health 7(2) (2010), 596–615.

[14] R. Vilalta, Y. Drissi, A perspective view and survey of meta-learning, J Artif Intell Rev 18(2) (2002), 77–95.