Empirical Bayes Model to Combine Signals of Adverse Drug Reactions

Rave Harpaz[1], William DuMouchel[2,3], Paea LePendu[1], and Nigam H. Shah[1]
[1]Center for Biomedical Informatics Research, Stanford University
[2]Oracle Health Sciences
[3]Observational Medical Outcomes Partnership

ABSTRACT

Data mining is a crucial tool for identifying risk signals of potential adverse drug reactions (ADRs). However, mining of ADR signals is currently limited to leveraging a single data source at a time. It is widely believed that combining ADR evidence from multiple data sources will result in a more accurate risk identification system. We present a methodology based on empirical Bayes modeling to combine ADR signals mined from $\sim$ 5 million adverse event reports collected by the FDA, and healthcare data corresponding to $\sim$ 46 million patients—the main two types of information sources currently employed for signal detection. Based on four sets of test cases (gold standard), we demonstrate that our method leads to a statistically significant and substantial improvement in signal detection accuracy, averaging 40% over the use of each source independently, and an area under the ROC curve of 0.87. We also compare the method with alternative supervised learning approaches, and argue that our approach is preferable as it does not require labeled (training) samples whose availability is currently limited. To our knowledge, this is the first effort to combine signals from these two complementary data sources, and to demonstrate the benefits of a computationally integrative strategy for drug safety surveillance.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences; H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms, Performance

Keywords

Empirical Bayes, signal detection, pharmacovigilance

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...$15.00.

1. INTRODUCTION

After a drug has been approved and is used on large, diverse populations, and for more varied periods of time, unanticipated adverse drug reactions (ADRs) may occur, which alter a drug's risk-benefit ratio enough to require remedial action. Post-approval ADRs are a major global health concern accounting for more than 2 million potentially preventable injuries, hospitalizations, and deaths each year in the US alone[15, 8], and associated costs estimated at $75 billion annually[6]. Pharmacovigilance, also known as drug safety surveillance, refers to the science and activities relating to the detection, assessment, understanding and prevention of ADRs in the post-approval period.

Data mining approaches that empower drug safety evaluators to analyze large volumes of data and to identify risk signals of potential ADRs, have proven to be a critical component in pharmacovigilance. Also known as signal detection methodologies, these data mining approaches are designed to compute measures of statistical association between pairs of drugs and clinical outcomes recorded in an underlying database. In a knowledge discovery scenario, the association statistics computed by data mining are interpreted as signal-scores, with larger values representing stronger associations, which are assumed more likely to represent true ADRs. Rankings of signal-scores or signal-score-thresholds are then used to flag associations worthy of further expert evaluation.

The US Food and Drug Administration (FDA) has maintained the Adverse Event Reporting System (AERS)[1] since 1968, which to date contains over 5 million spontaneous reports of suspected ADRs collected from healthcare professionals, consumers, and pharmaceutical companies. Each report includes one or more adverse events that appear to be associated with the administration of a drug, as well as indications and limited demographic information. Spontaneous reporting systems such as AERS communicate genuine health concerns, cover large populations, and are generally accessible for analysis. Since its inception AERS has supported regulatory decisions for a long list of marketed drugs[22]. Notwithstanding, AERS suffers from a range of recognized limitations including: reporting biases, misattribution of causality in reported ADRs, missing and incomplete data, duplicated reporting, and lack of true exposure information[7, 20].

Pharmacovigilance has predominantly relied on spontaneous reporting systems such as AERS. However, given their limitations, and the expanding availability and tremendous potential of healthcare data to advance pharmacovigilance, research efforts are now shifting towards the secondary use of large healthcare databases[10] such as electronic health records and administrative claims that typically contain: time-stamped interventions, procedures, diagnoses, medications, medical narratives, and billing codes. Unlike spontaneous reports, healthcare data reflect 'real-world' routine clinical care recorded over long periods of time. As such, they contain a more complete record of the patient's medical history, treatments, conditions, and potential risk factors.

US Food and Drug Administration Amendments Act of 2007 requires the FDA to develop a national system for monitoring medical product safety based on diverse healthcare data[2]. In 2008 the FDA launched the Sentinel Initiative[16] to meet this requirement. As part of this effort the Observational Medical Outcomes Partnership (OMOP)[19, 18, 4] was established to conduct methodological research to support the development of a national risk identification and analysis system, and a similar research initiative called the EU-ADR project was initiated in Europe[9].

The FDA routinely applies data mining to AERS in order to monitor and identify new safety signals of ADRs that warrant further attention. Similar surveillance strategies can be applied to healthcare data as demonstrated through pilot studies by the OMOP and the EU-ADR project. Although both AERS and healthcare data present unique challenges in its use, a common belief is that they may complement each other along several dimensions that will improve pharmacovigilance.

This paper presents a methodology to combine signals generated from spontaneous reports and healthcare data, and importantly aims to demonstrate that signal detection accuracy can be improved by such an integrative strategy. To our knowledge, this work is the first to explicitly and computationally combine signals from the two sources. We further argue that the proposed methodology is preferable to alternative approaches based on supervised learning that may be employed.

The proposed methodology draws parallels from statistical meta-analysis, and is based on empirical Bayes modeling where ADR signal-scores mined from each data source are modeled concomitantly using a Bayesian two-stage normal/normal model whose two hyper-parameters are estimated from the data using the expectation maximization (EM) algorithm. The output of the method is composite (combined) signal-scores consolidating the statistical evidence supplied by the source-dependent signal-scores. The methodology is applied to signals mined from $\sim$ 5 million public domain AERS reports, and healthcare data corresponding to $\sim$ 46 million patients captured in the 'OMOP results set'. The performance of the method (i.e., the signal detection accuracy of the combined signal-scores) is measured based on a validated gold standard created by the OMOP, totaling 380 positive and negative ADR test cases, and spanning four clinical outcomes.

The remainder of the paper is organized as follows: Section 2 provides background on signal detection methodologies and the specific methods used in this paper. Section 3 presents the proposed method to combine signals. Section 4 describes the experiments performed including data sources and the evaluation process. Section 5 provides the results and discussion, and Section 6 presents the conclusion. The term "signal-score" (or just signal) will be used to refer to a statistic that represents the strength of statistical association between a drug-outcome combination recorded in an underlying database.

Table 1: Contingency table used to compute association statistics for SRS-based signal detection

	reports w outcome	reports wo outcome
reports w drug	a	b
reports wo drug	c	d

Each cell contains report counts. Reports are assigned to cells based on whether they contain a specific drug and outcome. The table must be computed for every drug–outcome combination being considered.

2. MINING ADR SIGNALS

Due to inherent differences between spontaneous reporting systems (SRS) and healthcare databases, different data mining (signal detection) approaches are usually applied to each.

2.1 Mining Signals from SRS

Currently, the main driving force behind SRS-based signal detection is an approach referred to as *disproportionality analysis*, which aims to quantify the degree to which a reported drug-outcome combination co-occurs in the data "disproportionally" as compared with what would be expected if there were no statistical association between the drug and the outcome. All signal detection methodologies based on disproportionality analysis use the entries of Table 1 (or stratified versions thereof) to derive surrogate measures of statistical association. The table is computed for every drug-outcome combination being considered. The methodologies may differ with respect to the exact association measure that is used and the statistical adjustments that may be applied to the measure. The most widely cited measure is the relative reporting ratio (RRR), defined as the ratio between the number of reports in a SRS mentioning a specific drug-outcome combination to an expected number of reports under the assumption that the drug and outcome occur independently. The expected number of reports is calculated using all the reports in the SRS mentioning the drug or outcome as a proxy for the true value. Specifically, based on Table 1

$$\text{RRR} = \frac{(a+b+c+d)(a)}{(a+b)(a+c)}$$

where Pr(outcome|drug)/Pr(outcome) can be viewed as the probabilistic interpretation of RRR. A RRR=3, for example, would indicate that there are three times as many spontaneous reports mentioning a specific drug-outcome combination than would be expected by chance, which in turn may support the hypothesis of an ADR relationship between the drug and outcome. A true value of RRR close to 1 supports the hypothesis that there is no association between the drug and outcome.

The Multi-item Gamma Poisson Shrinker (MGPS)[12] used in this work to generate signals from AERS, is a leading SRS-based signal detection algorithm, which has been endorsed by the FDA as well as other regulatory agencies and pharmaceutical companies world-wide. MGPS computes a Bayesian regularized and stratified RRR designed to guard against false positive signals due to sampling variance, as well as account for biases due to temporal reporting trends and confounding by age and sex. It uses a Gamma-Poisson model to compute a centrality measure of the posterior distribution of the true RRR in the population. The measure is called empirical Bayes geometric mean (EBGM), and can be interpreted as the observed stratified RRR shrunken towards a prior when less data is available about the specific drug-outcome association being estimated. The prior is assumed to follow a bimodal Gamma distribution that models the RRRs of all distinct drug-outcome combinations in the SRS.

Drugs in AERS are entered verbatim, but are then usually mapped to their generic names or their active ingredients. Outcomes in AERS are coded using MedDRA[3] (a controlled vocabulary developed for ADR applications) usually at the 'preferred term' level of the MedDRA hierarchy. AERS currently includes about 4,000 mapped drugs, about 15,000 MedDRA preferred terms, and about 5 million distinct drug-outcome combinations appearing at least once that may be considered for analysis and for which associations statistics need to be computed. Signal detection methods such MGPS are usually applied to compute associations for all reported drug-outcome combinations in the SRS, typically on a quarterly basis whenever a new batch of spontaneous reports becomes available.

2.2 Mining Signals from Healthcare Data

Unlike SRS that are oriented towards pharmacovigilance, the secondary use of healthcare data requires special care to properly account for potential confounding biases (e.g., pre-existing risk factors) that may distort the estimation of a drug-outcome association. Although methods originally developed for SRS may be applied, an alternative class of approaches that are better equipped to deal with confounding is usually employed. This class of approaches is based on epidemiologic study designs that aim to control confounding by seeking to ensure that the two groups of subjects used to study an association (e.g., exposed/unexposed, cases/controls) are comparable with respect to potential confounding factors, which therefore cannot be the reason for an association. Another key distinguishing feature of this class of methods is their systematic use of temporal information, which is generally unavailable in SRS. The temporal information is used to establish various time frames, known as surveillance windows, drug/condition eras, or hazard periods, which are used to identify and count drug and outcome co-occurrences used in subsequent calculations, e.g., number of outcomes recorded within 30 days of drug exposure. Each method belonging to this class follows a different analytic paradigm and has multiple parameter settings corresponding to various study design choices, such as: length of surveillance window, type of comparator group, counting strategy, and confounding adjustment strategy.

Observational Screening (OS)[4] is a method developed by the OMOP that was used in this work to compute signals from healthcare data. Its preference over other methods will be made clear in Section 4.2 (greater signal detection accuracy). Under the parameter setting used in this work, OS represents a 'self-controlled' study design, wherein subjects serve as their own controls by comparing outcome rates for periods when a subject is exposed to a drug to periods when the subject is unexposed to the drug. Therefore, implicitly controlling for all time-invariant and subject-invariant confounders (e.g., comorbidities, smoking status, and chronic use of drugs) without the need for the confounders to be identified and measured. The core measure calculated by OS is a Screening Rate (SR) defined as

$$\text{SR} = \frac{\#\text{ of outcomes}}{\text{Total Time at Risk}}$$

The 'Time at Risk' for a drug is the length of exposure to the drug and an additional time period added to the end of the exposure (in this work Time at Risk=length of exposure+30 days). The 'Total Time at Risk' is an accumulated Time at Risk over all exposures. The number of outcomes counted towards SR is a count of outcomes occurring during the Time at Risk. The association statistic output by OS is a Screening Rate Ratio (SRR), defined as

$$\text{SRR} = \frac{\text{SR of exposed group}}{\text{SR of unexposed group}}$$

In this case (self-controlled design) the SR of the unexposed group is calculated by specifying a Time at Risk period prior to each drug exposure, whose length is the same as for the exposure period, and counting outcomes during that period. Fig. 1 provides an illustration of the components used in the calculation performed by OS for a specific drug-outcome pair and a population consisting of two patients. In this example 'SR of exposed group'=(1+1+2)/(2+3+5), 'SR of unexposed group'=(0+1+1)/(2+3+5), and SRR=(4/10)/(2/10)=2.

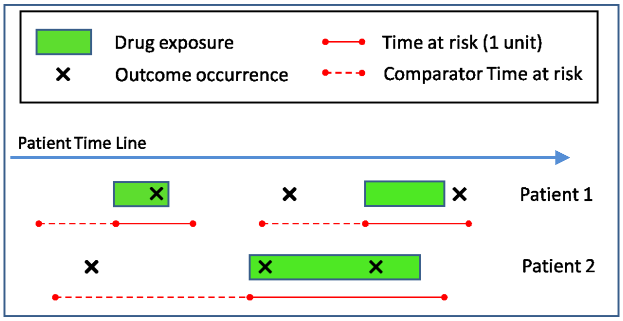

Figure 1: Illustration of the components used in the calculation performed by the Observation Screening signal detection method.

Regardless of the method or data source used, the output of a signal detection method for a specific drug-outcome pair is an association statistic (e.g., SRR, EBGM), and its lower and upper bound confidence interval (or alternatively its variance). The lower bound association statistic is often used as a signal-score instead of the point estimate as a suggested adjustment to reduce false signaling. A more detailed overview of pharmacovigilance data mining approaches can be found in ref. [14].

3. COMBINING ADR SIGNALS

3.1 Empirical Bayes

The empirical Bayes approach (EB) is often viewed as a compromise between classical (frequentist) and fully Bayesian approaches to statistical inference, borrowing ideas from each. EB starts by specifying a hierarchical model as do all Bayesian techniques. Hierarchical models in turn depend on a sequence of priors that must stop at some point with the remaining prior parameters assumed known. This is where EB separates from the fully Bayesian approach. Rather than assuming and specifying these final-stage prior parameters, the EB approach uses the observed data to estimate these parameters, and then proceeds as though they were known substituting them into the original Bayes quantities. This attribute was the main reason for employing EB in the current application. Namely, enjoying the flexibility and benefits of Bayesian modeling, but not having to specify prior parameters, which in the current application would have been based on nothing other than a "good guess".

3.2 The Bayes Model to Combine Signals

Suppose we have J pairs of signal-scores, where each pair corresponds to a unique drug-outcome association and each element in the pair corresponds to a signal originating from a different data source (e.g., AERS or healthcare data) that is computed by a possibly different signal detection method. The signals need not be based on the same statistic but are on approximately the same scale, and are assumed log-normally distributed.

Let y_{jk} denote the log of the j^{th} signal-score computed from the k^{th} data source, where $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Further, let $s_{jk}^2 = \text{Var}(y_{jk})$ be the accompanying observed variance of each signal-score, computed along with the signal. The goal is to estimate the unknown parameter μ_j denoting the combined (composite or pooled) signal-score of the j^{th} drug-outcome association, by observing $Y = \{ y_{jk} \}$ and $S = \{ s_{jk}^2 \}$. In the current application $K = 2$, but we use a more general formulation to emphasize that the framework allows for more than just two sources to be considered without a need to modify the methodology.

The two-stage hierarchical model describing Y and used to estimate μ_j is given by

$$y_{jk} \sim \text{N}(\mu_j, s_{jk}^2) \tag{1a}$$

$$\mu_j \sim \text{N}(\theta, \tau^2) \tag{1b}$$

where θ, τ^2 are hyper-parameters to be estimated from the data (discussed below). The model assumes that each tuple of observed signal-scores $(y_{j1}, \ldots, y_{jK})$ are random manifestations of normal process centered around the true but unknown combined signal-score μ_j, which itself is normally distributed (prior) around θ—a grand mean allowing for J related signals (e.g., signals related to the same clinical outcome) to borrow statistical support from one another.

Based on Eq. 1 the goal can be stated as computing the posterior distribution $\mu_j | Y, S, \theta, \tau^2$ and using $\text{E}[\mu_j | Y, S, \theta, \tau^2]$ as the estimate of μ_j. Further, according to Eq. 1 the joint density (data likelihood) is given by

$$f(Y, \boldsymbol{\mu} | \theta, \tau^2, S) = \prod_{j=1}^{J} \prod_{k=1}^{K} \text{N}(y_{jk} | \mu_j, s_{jk}^2) \prod_{j=1}^{J} \text{N}(\mu_j | \theta, \tau^2) \tag{2}$$

We denote by $\bar{y}_j$ a statistic that is meant to summarize the information (signal-scores) provided by each data source for a given drug-outcome association, and consider two common possibilities

$$\bar{y}_j = \sum_{k=1}^{K} y_{jk}/K \quad \text{with} \quad s_j^2 = \text{Var}(\bar{y}_j) = \sum_{k=1}^{K} s_{jk}^2/K^2 \tag{3a}$$

$$\bar{y}_j = \frac{\sum_{k=1}^{K} y_{jk}/s_{jk}^2}{\sum_{k=1}^{K} 1/s_{jk}^2} \quad \text{with} \quad s_j^2 = \text{Var}(\bar{y}_j) = \frac{\prod_k s_{jk}^2}{\sum_k s_{jk}^2} \tag{3b}$$

The summary statistic provided in Eq. 3(a) assumes that the information contributed by each data source is weighted equally, whereas 3(b) assumes that the contribution is proportional to the uncertainty (variance) associated with the signal-score supplied by each data source, i.e., more weight is assigned to a signal whose variance is smaller.

Having defined $\bar{y}_j$ (whether based on Eq. 3(a) or Eq. 3(b)) we approximate the joint density given in Eq. 2 by

$$f(Y, \boldsymbol{\mu} | \theta, \tau^2, S) = \prod_{j=1}^{J} \text{N}(\bar{y}_j | \mu_j, s_j^2) \text{N}(\mu_j | \theta, \tau^2) \tag{4}$$

Based on Eq. 4 it can be shown[13] that the posterior distribution of interest is

$$\mu_j | Y, S, \theta, \tau^2 \sim \text{N}(B_j \bar{y}_j + (1 - B_j)\theta, B_j s_j^2) \tag{5}$$

where

$$B_j = \frac{\tau^2}{\tau^2 + s_j^2}$$

Therefore, our estimate of μ_j is given by

$$\hat{\mu}_j = \text{E}[\mu_j | Y, S, \theta, \tau^2] = B_j \bar{y}_j + (1 - B_j)\theta \tag{6}$$

Eq. 6 shows that the estimated combined signal-score of the j^{th} drug-outcome association equals a weighted average of the prior mean θ and the statistic $\bar{y}_j$ summarizing the signal-scores $(y_{j1}, \ldots, y_{jK})$. The weights B_j and $(1 - B_j)$ are functions of the uncertainty (variance) associated with the two weighted extremes. That is, when the summary statistic $\bar{y}_j$ has smaller variance (s_j^2) more weight will be put on $\bar{y}_j$. Conversely, larger uncertainly will shrink $\bar{y}_j$ towards the prior mean θ . In this way, we are not only pooling associations across data sources, but are also allowing for the pooled signals μ_j to "borrow strength" from each other within similar groups to provide potentially more accurate estimates. This borrowing of strength is especially useful, as discussed later, when estimating ADR signals related to the same outcome or drug class.

3.3 Estimating the Hyper-parameters

To apply Eq. 6 we need to estimate the hyper-parameters θ and τ^2. Conditioned on μ_j the observations $\bar{y}_j$ are independently distributed. Based on this, the empirical Bayes approach uses the marginal likelihood in Eq. 7 to estimate the hyper-parameters θ, τ^2 of the model.

$$\begin{aligned}\prod_{j=1}^{J} f(\bar{y}_j | \theta, \tau^2, s_j^2) &= \prod_{j=1}^{J} \int \text{N}(\bar{y}_j | \mu_j, s_j^2) \text{N}(\mu_j | \theta, \tau^2) \text{d}\mu_j \\ &= \prod_{j=1}^{J} \text{N}(\bar{y}_j | \theta, \tau^2 + s_j^2) \end{aligned} \tag{7}$$

Because a closed form solution does not exist, we use the EM algorithm to obtain the maximum likelihood estimates of θ, τ^2 in Eq. 7. The EM algorithm, which in this situation offers a relatively simple alternative to other optimization techniques, can be applied by considering μ_j as a latent or missing variable, Eq. 7 as a missing-data likelihood, and Eq. 4 as the complete-data likelihood. Dempster et al. [11] showed that when the distribution of the complete-data (e.g., Eq. 4) belongs to an exponential family, or alternatively when the complete-data log-likelihood is linear in some sufficient statistic T (as in this case), then the E-step in the EM algorithm reduces to computing the posterior conditional expectation of T given the observed data, and the M-step reduces to substituting the expectation computed in the E-step in the expression for the complete-data maximum likelihood of the parameters to estimate. Therefore, given that $T_1 = \sum_j \mu_j$ and $T_2 = \sum_j \mu_j^2$ are sufficient statistics for θ and τ^2 respectively, and using t to index the current iteration, the EM steps can be stated as:

E-step:

$$T_1^{(t)} = \mathrm{E}\left[\sum_{j=1}^{J} \mu_j | Y, S, \theta^{(t)}, \tau^{(t)}\right] \tag{8a}$$
$$= \sum_{j=1}^{J} \left[B_j^{(t)} \bar{y}_j + (1 - B_j^{(t)})\theta^{(t)}\right]$$

$$T_2^{(t)} = \mathrm{E}\left[\sum_{j=1}^{J} \mu_j^2 | Y, S, \theta^{(t)}, \tau^{(t)}\right] \tag{8b}$$
$$= \sum_{j=1}^{J} \left[B_j^{(t)} \bar{y}_j + (1 - B_j^{(t)})\theta^{(t)}\right]^2 + \sum_{j=1}^{J} B_j^{(t)} s_j^2$$

M-step:

$$\theta^{(t+1)} = \frac{T_1^{(t)}}{J} \tag{9a}$$

$$\tau^{2^{(t+1)}} = \frac{T_2^{(t)}}{J} - \theta^{(t+1)^2} \tag{9b}$$

To summarize, the whole process for estimating the combined signal-scores $\hat{\mu}_j$, $j = 1, \ldots, J$, requires iteratively computing Eqs. 8–9 until convergence, and substituting the final estimates $\hat{\theta}, \hat{\tau}^2$ into Eq. 6. Crude estimates for θ and τ that can be used to seed/initialize the EM algorithm are given by:

$$\hat{\theta} = \frac{1}{J} \sum_{j=1}^{J} \bar{y}_j \tag{10a}$$

$$\hat{\tau}^2 = \frac{1}{J} \sum_{j=1}^{J} \left[(\bar{y}_j - \hat{\theta})^2 - s_j^2\right] \tag{10b}$$

Having estimated $\hat{\mu}_j$, lower bound signal-scores can be computed using the posterior variance $B_j s_j^2$ given in Eq. 5, e.g., $\hat{\mu}_j - Z_\alpha \sqrt{B_j s_j^2}$.

Table 2: Distribution of OMOP test cases used in the evaluation

Outcome	Positive Cases	Negative Cases	Total
Acute Renal Failure	22	58	80
Upper GI Bleed	24	66	90
Acute Liver Injury	77	36	113
Acute Myocardial Infarction	34	63	97
Total	157	223	380

4. EMPIRICAL ASSESSMENT

4.1 Gold Standard

The proposed methodology was evaluated on the basis of its ability to correctly signal (classify) a total of 380 positive and negative test cases (drug-outcome pairs), which are part of a gold standard created and thoroughly validated by the OMOP, and for which both AERS and healthcare data was available. Positive test cases are true ADR association asserted from drug labeling (mention of an outcome as an adverse reaction) and/or prior published research suggesting an association. Conversely, negative test cases are associations that lack this level of evidence in their labeling or the literature. The entire gold standard includes 181 drugs and is divided into four sets each associated with a unique outcome—**acute myocardial infarction**, **acute renal failure**, **acute liver injury**, and **upper gastrointestinal bleeding**, which represent four of the most significant and actively monitored drug safety outcomes[21]. A cross tabulation of the test cases by outcome is provided in Table 2.

4.2 Healthcare Data

The 'OMOP results set' is based on five medical databases comprised of administrative claims and electronic health records, which reflect the healthcare experience of about 74 million patients. To each of these databases the OMOP applied seven unique and commonly used methods to compute signal-scores of each drug-outcome pair in their gold standard. Each method follows a different analytic paradigm and has multiple parameter settings, as described in Section 2.2. In total, the OMOP results set contains ∼6 million signal-scores (and associated statistics) representing every combination of database, method, parameter setting, and drug-event pair in the gold standard. The result set is publicly available at http://omop.fnih.org/research.

As the set of signal-scores used in our evaluation, we selected from the OMOP results set signal-scores corresponding to the largest database—'MarketScan Commercial Claims and Encounters', which contains claims data corresponding ∼46 million patients and is abbreviated–CCAE. The signals were computed by the OS method (parameter setting referenced by Analysis-ID: 403002), which uniformly provided the best diagnostic accuracy across the four outcomes given data from CCAE, and explains the preference of the OS method over other methods (Section 2.2).

4.3 AERS Data

Using the public-release version of AERS we extracted a total of 4,784,337 spontaneous reports covering the period from 1968 through 2011Q3. The data was preprocessed (as suggested in the literature) to remove duplicate reports and correct terminological errors. To facilitate interoperability of terms and definitions used to describe drugs and outcomes in AERS and the OMOP results set, we mapped drug names in AERS to their ingredient level specification. MedDRA preferred terms used to specify outcomes in AERS were mapped to the four outcomes in the gold standard using broad MedDRA group definitions supplied by OMOP. The preprocessed AERS data was then loaded into the Empirica Signal V7.3 system (ESS)—a drug safety data mining application from Oracle Health Sciences[5].

Within ESS we applied the Multi-item Gamma Poisson Shrinker (Section 2.1) based on its standard parameter settings to generate signal-scores for each of the 380 tests cases in the gold standard.

4.4 Evaluation

Given pre-computed AERS and healthcare signal-scores corresponding to each of the 380 test cases, the proposed methodology to combine the signal-scores was applied independently to each of the four outcome sets of test cases. That is, each outcome was modeled separately—assuming that signals associated with the same outcome are statistically related and can therefore borrow support from each other, but are unrelated to signals associated with other outcomes and should not borrow support from them.

Based on OMOP's set of test cases, the performance (signal detection accuracy) of the resulting system (combined signal-scores) was compared against the performance of signals generated by each data source independently. Performance was measured using the threshold-independent measure—area under the receiver operating characteristic (ROC) curve (AUC), which is the most widely used index for measuring diagnostic accuracy.

The methodology was also compared against linear and non-linear supervised classification/prediction algorithms as potential competing approaches to combine signals. These can be applied by treating the signal-scores y_1, y_2 as two features/predictors, interpreting the decision/predicted values as combined signal-scores, and by using subsets of the OMOP test cases as training and testing samples. In this application, a linear classifier will have the form $f(w_0 + w_1 y_1 + w_2 y_2)$, where f is a strictly monotonic function that maps a linear combination of signal-scores to a decision/predicted value. Since a ROC curve is invariant to monotonic transformations,

$$\begin{aligned} &\text{AUC}\left(f(w_0 + w_1 y_1 + w_2 y_2)\right) \\ =&\text{AUC}\left(w_0 + w_1 y_1 + w_2 y_2\right) \\ =&\text{AUC}\left(w_1 y_1 + w_2 y_2\right) \\ =&\text{AUC}\left(y_1 + w y_2\right) \end{aligned}$$

Therefore,

$$\max_{-\infty < w < \infty} \text{AUC}\left(y_1 + w y_2\right) \tag{11}$$

is an upper bound to the AUC attainable by any specific linear classifier. So instead of evaluating a specific set of linear classifiers (e.g., logistic regression, LDA, perceptron, linear SVM) we cast our evaluation to computing the AUC in Eq. 11, and refer to the hypothetical method producing this AUC as the "Optimal Linear Classifier". The maximization of Eq. 11 was performed using a 1D grid search (intervals=0.01, -10<w<10). Because the same generalization does not hold for non-linear classifiers we used the radial basis kernel SVM (in the e1701 R package) as a representative. Both approaches were evaluated using 5-fold cross-validation based on class-stratified samples. The results were averaged to produce a single AUC.

5. RESULTS & DISCUSSION

The main results of our evaluation are summarized in Table 3, which displays a comparison of AUC-based signal detection accuracy across the data sources/methods evaluated. We define 'Relative Improvement' as

$$\frac{\text{AUC(Combined)} - \max(\text{AUC(AERS)}, \text{AUC(Healthcare)})}{1 - \max(\text{AUC(AERS)}, \text{AUC(Healthcare)})} \tag{12}$$

i.e, the proportion of error reduction gained by using the combined signal-scores over the better performing individual data source signal-scores.

Overall, Table 3 demonstrates that combining signals across AERS and healthcare data using the proposed methodology leads to an overall substantial improvement. The results also demonstrate that the improvement is replicated across analysis of different outcomes. Since the method is unlikely to transform two strong signals into a weak signal, nor is it expected to transform two weak signals into a strong composite signal, the success of the method can be linked to cases where the two data sources provide inconsistent or conflicting statistical information that is resolved by the method's ability to consolidate statistical information. The table also shows that the performance of signal detection varies across outcomes—supporting the design and modeling decision to treat each outcome separately. The relative improvement ranges from 20% for the outcome–acute myocardial infarction, to an improvement of 56% for acute renal failure, with an average improvement of 40%. Signal detection accuracy ranges from AUC=0.76 for acute myocardial infarction to AUC=0.94 for acute renal failure, with an average AUC=0.87—a level of accuracy considered sufficient in other widely used clinical diagnostic tests (e.g., prostate cancer, and breast cancer)[18]. Similar performance patterns were observed (not displayed) when using the lower bound signal-scores (Section 2), with performance ranging from AUC=0.78 (myocardial infarction) to AUC=0.96 (renal failure), and an improvement ranging from 9% (myocardial infarction) to 62% (acute renal failure).

The summary statistic $\bar{y}_j$ underlying the results displayed in Table 3 is the one proposed in Eq. 3(b) (inverse-variance weighting), as it resulted in greater signal detection accuracy than the alternative possibility (equal weighting of the data sources), which averaged a 35% improvement over each data source . We note however that this type of weighting (Eq. 3(b)) could pose a problem when one of the data sources being considered is much larger than the others, in which case it may dominate the weighting for certain associations. We also note that using $\hat{\mu}_j$ as the combined signal-score resulted in greater accuracy by an average of 5% over using just $\bar{y}_j$ (also a possibility), showing that the modeling approach, the notion of Bayesian shrinkage and that of allowing signals to borrow support from each other, is beneficial.

To test if the improvements were statistically significant we computed a one-sided p-value for the hypothesis that the difference between the performance (AUC) of the combined signals and those of the individual data sources was is

Table 3: Comparison of signal detection accuracy based on AUC

Outcome	AERS	Healthcare	Combined	Relative Improvement	Optimal Linear Classifier	Non-Linear SVM
Acute Renal Failure	0.86	0.81	0.94	**56%**	0.96	0.95
Upper GI Bleed	0.89	0.73	0.94	**49%**	0.95	0.86
Acute Liver Injury	0.70	0.76	0.85	**37%**	0.83	0.76
Acute Myocardial Infarction	0.64	0.70	0.76	**20%**	0.68	0.64
Average	0.77	0.75	0.87	**40%**	0.85	0.81

AERS, Healthcare: AUC of signal-scores generated from AERS and healthcare data independently. Combined: AUC of the combined signal-scores generated using the proposed method. Last two columns: AUC of two potentially competing methods to generate combined signal-scores.

greater than 0. The tests were computed using the R package pROC[17], which uses a non-parametric test for correlated ROCs or bootstrapping. To ensure the p-values were computed based on a large enough sample of signal-scores, and to get a single answer representing all outcomes, we pooled the signal associated with each outcome into a single set of signal-scores, producing three sets of signals-scores for the three signal detection approaches that were tested against each other. The p-value for the difference between the combined signals and those of AERS was 0.001, and for healthcare was 5.9e-08, demonstrating that the improvements were statistically significant at the standard levels commonly used (e.g. p-value<0.05).

Given that there are currently no pharmacovigilance guidelines recommending appropriate thresholds or appropriate sensitivity-specificity tradeoffs, we do not provide threshold-dependent performance metrics, as those would currently carry little value. Nonetheless, the information provided through Fig. 2—a comparison of ROC curves—may be used as a substitute from which point-wise performance values may be extracted. Importantly, Fig. 2 depicts a general pattern of containment between the ROC curves of the combined signal-scores and those of the individual sources (with the exclusion of acute myocardial infarction), suggesting that the combined signal-scores provide greater accuracy at any single point of sensitivity, specificity, or signal-threshold that may be chosen in practice. It appears that for the case of myocardial infarction a lower tolerance for false positives will lead at a certain point to no improvement over the healthcare-based signal-scores. However, for the region of false positive rates likely to be tolerated in practice (discussed next) the performance of the combined signal-scores for acute myocardial infarction is still greater.

The partial-AUC is often used as an alternative measure to the full AUC when the goal is to consider only certain ranges of sensitivity or specificity which are deemed clinically relevant. The partial-AUC is simply the area under a portion of the ROC curve, often defined as the area between two false positive rates. A partial-AUC at 0.3 false positive rate (PAUC30), i.e., partial-AUC when specificity>0.7, has been previously suggested as a potential region of clinical relevancy for signal detection assessment[18]. Table 4 provides a comparison of signal detection accuracy among the individual and combined signal-scores based on PAUC30. The table shows that also in this restricted ROC space the proposed method provides greater accuracy across all outcomes, and similar levels of improvement, averaging 33%. The relative improvement is defined as for the full AUC, but in this case the largest possible AUC is 0.3 so the denominator in Eq. 12 is changed accordingly (0.3 replaces the value 1). The p-values for the partial-AUC improvement (computed as for the full AUC) over AERS and healthcare were 0.001 and 2.171e-09 respectively.

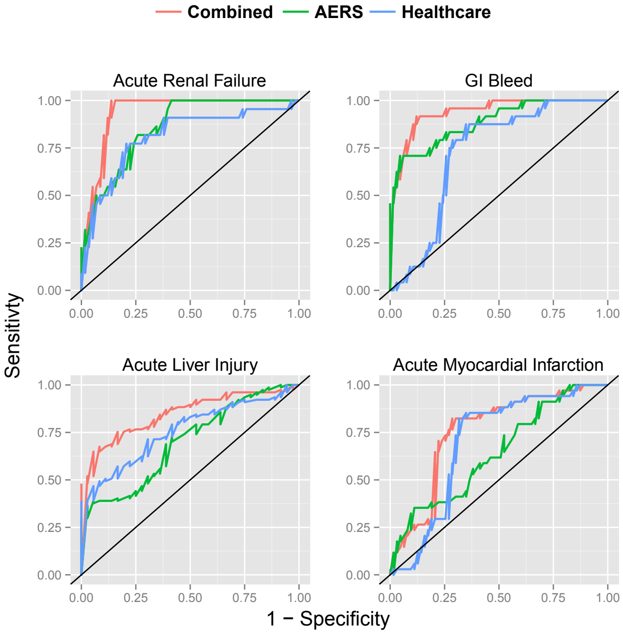

Figure 2: ROC curves of signal-scores generated from AERS, healthcare data, and the combined methodology.

Another important finding demonstrated through Table 3 is that the performance of the proposed methodology is comparable and on average slightly better than the potentially competing classification/prediction algorithms evaluated, which unlike the proposed methodology require labeled examples to train (fit) a model. Given the difficulty associated with identifying large sets of drug-outcome pairs with validated causative relationships that would be necessary to apply these alternative methods, we argue that approaches such as the proposed, that may be perceived as unsupervised learning, would be advantageous.

Similar to other applications that require optimization, the success of the proposed methodology depends on the ability of the EM algorithm to identify the correct solu-

Table 4: Comparison of signal detection accuracy based on partial-AUC at 0.3 false positive rate

Outcome	AERS	Healthcare	Combined	Relative Improv.
ARF	0.17	0.17	0.24	**51%**
GIB	0.22	0.07	0.25	**39%**
ALI	0.12	0.16	0.20	**34%**
AMI	0.09	0.06	0.11	**10%**
Avg.	0.15	0.11	0.20	**33%**

ARF: acute renal failure, GIB: upper gastrointestinal bleeding ALI: acute liver failure, AMI: acute myocardial infarction.

tions (hyper-parameters of the model) and on its convergence properties. Fig. 3 displays the solution spaces (surface plots) of the hyper-parameters θ, τ to be estimated by the EM algorithm for each outcome. The figure suggests that the solution space is concave across all outcomes, and therefore that the solutions identified by the EM algorithm should correspond to a global maxima, and also should not be sensitive to the initial EM values (the concavity of the solution space is data dependent and cannot be proved analytically). Using arbitrary initial EM values set to $\theta = 0$, $\tau = 1$, Fig. 4 demonstrates that convergence is rapid, with at most 11 steps required to arrive at a solution (convergence tolerance=10^{-9}). Although convergence does not appear to be an issue with the current data, using the initial EM values suggested in Eq. 10 resulted in a faster convergence by a factor of almost 2.

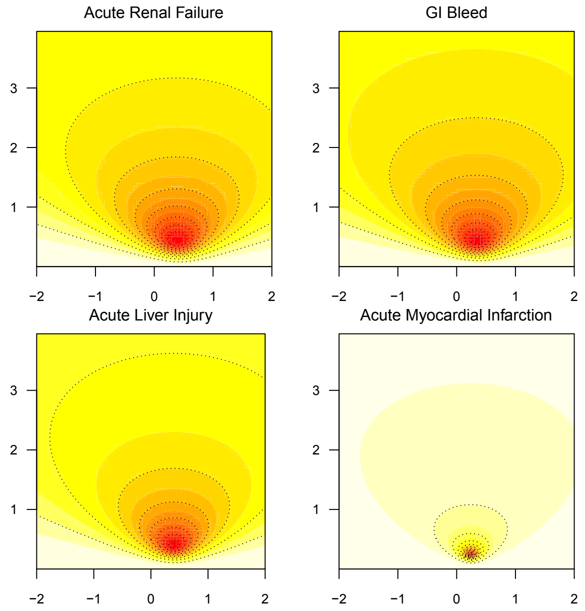

Figure 3: Surface plots of the hyper-parameters' (θ, τ) solution space, estimated by the EM algorithm. The color pallet and contours reflect varying values of log likelihood with 'hotter' colors corresponding to larger likelihood.

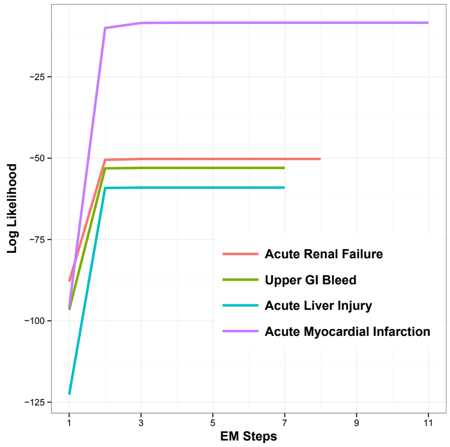

Figure 4: Convergence of the EM algorithm used to estimate the hyper-parameters θ, τ for each outcome.

6. CONCLUSION

The synthesis of evidence from multiple streams of information has been an integral part of pharmacovigilance. Yet, it is currently carried out by human experts on an ad hoc basis, in a rather qualitative manner, and usually after a signal is generated. Most signal detection strategies are currently based on data associated with a single source. Given the relative maturity of surveillance based on spontaneous reporting, the recent progress made in the use of healthcare data, and the expectation that the two sources may complement each other along different dimensions, it appears that the time is ripe to consider computational approaches to combine information from these two types of information sources and possibly other sources.

This paper presents the first effort to explicitly and computationally combine ADR signals from spontaneous reporting systems and healthcare data to improve the accuracy of uninformed hypothesis-free signal detection. Improving the accuracy of ADR signal detection is paramount to data mining for pharamcovigilance.

The methodology was applied to signals generated by established methods from two large databases of high quality, and is evaluated using a large thoroughly validated gold standard; thus minimizing concerns related to the reliability of data and resources used in this work. Through different analyses we demonstrated that the proposed methodology leads to a statistically significant and substantial improvement of signal detection accuracy over the use of each source independently. We also showed that the improvement is replicated over analysis of different outcomes, and therefore may generalize to other clinical outcomes. The methodology is relatively simple and efficient to compute, and is generalizable to the inclusion of additional data sources with no modification. Its performance was shown to be comparable to alternative approaches based on supervised learning, which unlike our approach, have the limitation of requiring labeled training samples, whose availability is limited. The methodology can be used to analyze specific outcomes, or it can be used as an add-on to routine data mining runs currently performed by various organizations that have access to both types of data sources, e.g., the FDA through the sentinel network. The public availability of SRS such as AERS also make this methodology available to a wider range of entities that specialize in, and process, large quantities of healthcare data.

Finally, it is possible that the use of different combinations of data and signal generation algorithms will lead to different performance characteristics or possibly a different conclusion. Therefore further research is needed to fully understand the dependence of the performance, and the success of the paradigm, on the variety of data and methods that can be used. Likewise, additional research is required to investigate methodological extensions and the potential inclusion of other healthcare databases into the framework.

7. ACKNOWLEDGMENTS

This research was supported in part by NIH grant U54-HG004028 for the National Center for Biomedical Ontology, the FDA's ORISE Fellowship and by the Department of Medicine at Stanford University. We extend our gratitude to Oracle's Health Sciences Division for supplying us with the AERS data and analysis software, and to David Madigan and Patrick Ryan from the OMOP for providing constructive suggestions and guidance on the use of the OMOP results set.

8. REFERENCES

[1] Adverse Event Reporting System. http://www.fda.gov/cder/aers/default.htm.

[2] Food and Drug Administration Amendments Act (FDAAA) of 2007.

[3] Medical Dictionary for Regulatory Activities (MedDRA). http://www.meddramsso.com/.

[4] Observational Medical Outcomes Partnership (OMOP). http://omop.fnih.org/.

[5] Oracle Health Sciences. http://www.oracle.com/us/products/applications/health-sciences/safety/empirica-signal/index.html.

[6] S. Ahmad. Adverse drug event monitoring at the food and drug administration - your report can make a difference. *Journal of General Internal Medicine*, 18(1):57–60, 2003.

[7] A. Bate and S. Evans. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol.Drug Saf*, 18(6):427–436, 2009.

[8] D. Classen, S. Pestotnik, R. Evans, J. Lloyd, and J. Burke. Adverse drug events in hospitalized patients. excess length of stay, extra costs, and attributable mortality. *JAMA*, 277(4):301–306, 1997.

[9] P. Coloma, M. Schuemie, G. Trifiro, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, C. Giaquinto, G. Corrao, L. Pedersen, L. J. van der, and M. Sturkenboom. Combining electronic healthcare databases in europe to allow for large-scale drug safety monitoring: the EU-ADR project. *Pharmacoepidemiol.Drug Saf*, 20(1):1–11, 2011.

[10] P. M. Coloma, G. Trifiro, V. Patadia, and M. Sturkenboom. Postmarketing safety surveillance. *Drug Safety*, pages 1–15, 2013.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[12] W. DuMouchel. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat.*, 53(3):177–190, 1999.

[13] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, Second Edition.* Chapman and Hall/CRC, 2003.

[14] 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. *Nature-Clin Pharmacol Ther*, 91(6):1010–1021, 2012.

[15] J. Lazarou, B. Pomeranz, and P. Corey. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*, 279(15):1200–1205, 1998.

[16] R. Platt, M. Wilson, K. Chan, J. Benner, J. Marchibroda, and M. McClellan. The new Sentinel Network - improving the evidence of medical-product safety. *New England Journal of Medicine*, 361(7):645–647, 2009.

[17] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Muller. pROC: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12(1):77, 2011.

[18] P. B. Ryan, D. Madigan, P. E. Stang, J. Marc Overhage, J. A. Racoosin, and A. G. Hartzema. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Statistics in Medicine*, 31(30):4401–4415, 2012.

[19] P. Stang, P. Ryan, J. Racoosin, J. Overhage, A. 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. *Annals of Internal Medicine*, 153(9):600–W206, 2010.

[20] W. Stephenson and M. Hauben. Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiol.Drug Saf*, 16(4):359–365, 2007.

[21] G. Trifiro, A. Pariente, P. M. Coloma, J. A. Kors, G. Polimeni, G. Miremont-Salame, M. A. Catania, F. Salvo, A. David, N. Moore, A. P. Caputi, M. Sturkenboom, M. Molokhia, J. Hippisley-Cox, C. D. Acedo, J. van der Lei, and A. Fourrier-Reglat. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiology and Drug Safety*, 18(12):1176–1184, 2009.

[22] D. Wysowski and L. Swartz. Adverse drug event surveillance and drug withdrawals in the united states, 1969-2002 - the importance of reporting suspected reactions. *Archives of Internal Medicine*, 165(12):1363–1369, 2005.