Toward multimodal signal detection of adverse drug reactions

Rave Harpaza, William DuMouchel, Martijn Schuemie, Olivier Bodenreider, Carol Friedman, Eric Horvitz, Anna Ripple, Alfred Sorbello, Ryen W. White, Rainer Winnenburg, Nigam H. Shah

a Oracle Health Sciences, Bedford, MA, United States
b Janssen Research and Development, Titusville, NJ, United States
c National Library of Medicine, NIH, Bethesda, MD, United States
d Columbia University, New York, NY, United States
e Microsoft Research, Redmond, WA, United States
f U.S. FDA, Silver Spring, MD, United States
g Stanford University, Stanford, CA, United States

ARTICLE INFO

Keywords:
Drug safety
Pharmacovigilance
Adverse drug reactions
Signal detection
Multimodal

ABSTRACT

Objective: Improving mechanisms to detect adverse drug reactions (ADRs) is key to strengthening post-marketing drug safety surveillance. Signal detection is presently unimodal, relying on a single information source. Multimodal signal detection is based on jointly analyzing multiple information sources. Building on, and expanding the work done in prior studies, the aim of the article is to further research on multimodal signal detection, explore its potential benefits, and propose methods for its construction and evaluation.

Material and methods: Four data sources are investigated; FDA’s adverse event reporting system, insurance claims, the MEDLINE citation database, and the logs of major Web search engines. Published methods are used to generate and combine signals from each data source. Two distinct reference benchmarks corresponding to well-established and recently labeled ADRs respectively are used to evaluate the performance of multimodal signal detection in terms of area under the ROC curve (AUC) and lead-time-to-detection, with the latter relative to labeling revision dates.

Results: Limited to our reference benchmarks, multimodal signal detection provides AUC improvements ranging from 0.04 to 0.09 based on a widely used evaluation benchmark, and a comparative added lead-time of 7–22 months relative to labeling revision dates.

Conclusions: The results support the notion that utilizing and jointly analyzing multiple data sources may lead to improved signal detection. Given certain data and benchmark limitations, the early stage of development, and the complexity of ADRs, it is currently not possible to make definitive statements about the ultimate utility of the concept. Continued development of multimodal signal detection requires a deeper understanding the data sources used, additional benchmarks, and further research on methods to generate and synthesize signals.

1. Introduction

The increasing harm and monetary burden associated with adverse drug reactions (ADRs) has made post-marketing drug safety surveillance (DSS) a top priority for health systems worldwide, and led to new legislative and research initiatives [1–6]. DSS presently relies on analysis of adverse event reports stored in spontaneous reporting systems (SRS) [7,8] such as the U.S. Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) [9,10] and the VigiBase maintained by the World Health Organization (WHO) [11].

The complexity associated with numerous natural phenomena (such as the mechanisms, occurrence, or identification of ADRs) makes it unlikely that a single information source provides complete knowledge of the phenomenon of interest. In many scientific domains, information about a given phenomenon can be acquired from multiple sources. Each such acquisition framework is referred to as a modality and is associated with one data source [12]. A system that provides access to information from multiple modalities is known as multimodal. The fusion and joint analysis of multiple modalities promises a more unified and global view of the problem at hand, and a solution that is greater than the sum of its parts.

The benefits that may come from multimodality have recently been
extended to DSS with the ability to process and analyze new kinds of observational, experimental, and knowledge-based data sources containing pertinent information about the effects, mechanisms, and safety of medical products. Among these data sources are: electronic health records and administrative claims [6,13–17], the biomedical literature [18–20], social media (e.g., health forums and social networks) [21–24], behavioral data drawn from the logs of search engines [25–27], and mechanistic information extracted from chemical and biological knowledge bases [28]. Together with SRS, each data source promises to provide a unique vantage point for shaping our understanding of a drug’s safety profile, for contributing scientific evidence needed for causal assessment of adverse reactions, and for improving the timeliness of ADR identification.

In recent years, a large amount of research has been conducted on repurposing and analyzing safety information from such data sources, but relatively less progress has been made on development of multimodal approaches to jointly analyze the information provided by each data source.

A core component of DSS is ADR detection, and use of computational techniques known as signal detection are among its primary facilitators. These techniques enable drug safety researchers to analyze large volumes of data and generate hypotheses (signals) of new post-approval ADRs. Upon review, strong signals may lead to regulatory interventions such as drug withdrawals and the issuance of public warnings. From a technical standpoint, signal detection consists of applying computational methodologies to large databases in order to identify unexpected associations between drugs and potential adverse reactions [10].

Prior studies have demonstrated that pair-wise combinations of safety signals from several sources with FAERS can improve the accuracy of signal detection. Examples of signals that have been combined with those from FAERS include signals from EHRs [29,30], claims data [31], biomedical literature [32], chemical data [33], and Internet search logs [25]. Additional studies with less focus on signaling considered other forms of joint analysis of safety data [34,35].

The aim of this article is to strengthen empirical support for multimodal signal detection and provide additional insights on its performance characteristics. Building on, and extending the work performed in prior studies, we consider the joint analysis of more than just two data sources, and investigate a larger number of methods to synthesize signal statistics. Rather than relying on a single benchmark of well-known ADRs, we extend the evaluation with an additional benchmark of relatively new ADRs, and importantly, we incorporate the time aspect of signal detection into our performance evaluation. In doing so we propose new performance indices that blend both the accuracy of detection and the lead-time-to-detection.

We envision a multimodal signal detection system as one which pools and aggregates signal statistics from multiple data sources to produce a composite signal statistic (Fig. 1). In characterizing such a system, we examine four data sources, including three representative sources that would likely play a significant role in future systems, and a fourth novel, yet non-standard source that demonstrates promise. We consider other forms of joint analysis of safety data [34,35].

The FAERS data consists of approximately six million reports collected by the FDA since the inception of the system (1968) to 2014Q2 [37]. The reports were preprocessed to correct reporting artifacts and were made available by Oracle Health Sciences.

Analysis of healthcare data was made possible through the Innovation in Medical Evidence Development and Surveillance Research Laboratory [38], which provides a secure computing environment, research tools, and five de-identified healthcare datasets for analysis. Of these datasets, we used the largest dataset called ‘MarketScan Commercial Claims and Encounters’, containing administrative claims from a privately insured U.S. population of approximately 142 million patients spanning the years 2003–2013 (inclusive).

Analysis of NLM’s Medical Subject Headings (MeSH®) [39] descriptors was used to generate signals from MEDLINE bibliographic citations spanning the period from 1968 to May 2014. The algorithms for computationally identifying and extracting MEDLINE MeSH terms associated with ADRs are described by Winnenburg et al. [20]. The process resulted in approximately 360,000 ADR-related article citations and approximately 500,000 unique drug-event pairs potentially associated as ADRs, which were supplied along with other article metadata by the NLM.

De-identified search queries submitted to the Google, Bing, and Yahoo search engines by 80 million users during the period 2011–2013 were analyzed to generate signals based on behavioral data. This data source was accessed in collaboration with Microsoft Research, and described by White et al. [25].

Our retrospective performance evaluation uses the Observational Medical Outcomes Partnership (OMOP) benchmark [40]. The OMOP benchmark consists of 399 unique drug-event pairs (test cases) of which 165 represent established true ADRs and 234 represent negative controls—pairs that are highly unlikely to be associated. Each pair is made of one of 181 drugs, and one of four serious adverse events (acute myocardial infarction, acute renal failure, acute liver injury, and upper

![Fig. 1. Architecture of a multimodal signal detection system. Signal statistics are pooled and aggregated from different data sources to produce a composite signal statistic.](image-url)
The description of signal statistics (observed to expected ratios) generated from each data source.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Parameters</th>
<th>Method description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAERS</td>
<td>( c = 1, \alpha = 0.5 )</td>
<td>( N ) corresponds to number of spontaneous reports containing both the drug and event investigated. ( E ) corresponds to a stratified expected count under the assumption of independence (providing some level of confounding control). The stratification is done by age (9 categories), gender, and year of report. The method produces a disproportionality statistic equivalent to exponential ((2)) of the IC statistic developed by the WHO [48], and thus is fundamentally the same method. The widely applied MGPS [49] method provided essentially equivalent performance (for this study), but its calculation diverges from the consistent formulation chosen for this study.</td>
</tr>
<tr>
<td>Claims</td>
<td>( c = \frac{t_a}{t_b}, \alpha = 0.5 )</td>
<td>SCC [50] 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, thus providing a mechanism to account for time-invariant confounding factors. In SCC ( N ) represents the total number of outcomes (events) recorded for patients in a database, which occur within pre-defined exposure windows (time at risk), and ( t_b ) is the sum of the number of days in these exposure windows. ( E ) and ( t_b ) are the equivalent quantities for the comparator group (same patients for unexposed periods). Averaged across the four OMOP events, SCC was found to be the best performing method [51].</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>( c = 1, \alpha = 0.5 )</td>
<td>( N ) represents the number of articles including MeSH terms for the drug and event investigated, and ( E ) is its standard expected value under the independence assumption. No confounding control was applied</td>
</tr>
<tr>
<td>Web Logs</td>
<td>( c = \frac{t_a}{t_b}, \alpha = 0.5 )</td>
<td>Method based on the SRR statistic described by White et al. [25], which is produced by a method similar to SCC. ( N ) represents the number of unique users that queried for the drug and event of interest within a pre-defined time period ( t_a ) after the first query for the drug of interest, and ( E ) and ( t_b ) are the equivalent quantities prior to the first query for the drug of interest. ( t_a ) and ( t_b ) were both set to 60 days.</td>
</tr>
</tbody>
</table>

The descriptions refer to the observed to expected ratio \( s = \frac{(N + \alpha)(cE + \alpha)}{cN + \alpha} \) defined in Eq. (1).

As our core signal statistic we use a Bayesian smoothed observed-to-expected ratio. This signal statistic is computed for each combination of test case, analysis period, and data source being evaluated, and is defined as

\[
s = \frac{N + \alpha}{cE + \alpha}
\]  

(1)

where \( N \) denotes the observed co-occurrence rate of a particular drug-event pair (test case), \( E \) its expected rate (under certain conditions), \( c \) a normalizing factor, and \( \alpha \) a smoothing parameter. This ratio corresponds to the posterior expectation of a Gamma-Poisson model of counts, and forms the basis of several routinely applied Bayesian signal detection algorithms for SRS including those of both the WHO [47,48], and the FDA [9,10,49]. For example, the commonly cited IC statistic [48] originally proposed by the WHO is equivalent to the log2 of this ratio with \( c = 1, \alpha = 0.5 \). The Bayesian nature of this ratio provides the benefits of smoothing (shrinkage) and imputation.

Signal statistics for FAERS and MEDLINE were generated based on Eq. (1) directly. For our claims and search logs data signals were generated by the self-controlled cohort (SCC) method developed by the OMOP [50], and the method described by White et al. [25] respectively. The signals produced by these two methods were then cast to the ratio of Eq. (1), thus enabling the use of a uniform signal statistic across all data sources. The suggested value \( \alpha = 0.5 \) was used throughout our study.

When no data is observed for the particular test case \( N = 0 \), it can be seen from Eq. (1) that a signal statistic is imputed for the test case. This allowed the large majority of test cases to be included in our evaluation rather than being unused or misused with trivial values. Test cases whose underlying mapped drug or event were not captured in any of the data sources \( E = 0 \) at a given analysis period (e.g., drug approved after the analysis period) were omitted from the analysis period.

For each evaluation category we used the better performing of either the ratio above or the lower 5% limit of its credible interval often used a ‘safer’ alternative [10]. The lower 5% limit of the credible interval is computed using the quantile function of the Gamma distribution (posterior distribution of the Gamma-Poisson model underlying the signal statistics) with shape and rate parameters \( N + \alpha \) and \( cE + \alpha \) respectively.

2.4. Combining signal statistics

To generate composite signal statistics, we examined two classes of methodologies: those based on predictive algorithms, and those based on different weighted averaging schemes.

The predictive algorithms we investigated included: random forests, logistic regression, support vectors machines, and a methodology we call naive Bayes smoothing developed for this study. The predictors/features used by these methods are the signal statistics \( s \) (Eq. (1)) generated from each data source at the analysis period. The composite signal score produced by these methods represents the probability that a given association is true given the predictors’ values.

The class of methods based on weighted averaging included the arithmetic and geometric averages, the weighted averaging implied by fixed and random effects meta-analysis, and the averaging implied by an earlier method we developed [31]. The averaging was composed of the signal statistics \( s \) (Eq. (1)) generated from each data sources together with their standard errors. Table 2 provides additional details about the methodologies we evaluated.

gastrointestinal bleeding). 40 potentially misclassified negative controls were omitted from the benchmark as recommended by Hauben et al. [41]. The OMOP benchmark was designed for retrospective evaluations, in which all data available from a given data source is typically used to evaluate historical associations.

Our backdated prospective performance evaluation uses the time-indexed benchmark [42], which is derived from drug labeling revisions issued by the FDA throughout the year 2013. The benchmark includes 62 positive test cases that represent recently labeled ADRs, and 75 negative controls. It covers 44 drugs and 38 adverse events ranging from mild to rare and serious. Each positive test case is time-stamped and had negative controls. It covers 44 drugs and 38 adverse events ranging from mild to rare and serious. Each positive test case is time-stamped and is de-
Table 2

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random forests</td>
<td>Uses the R package <code>randomForest</code>. <strong>Predictors:</strong> s_1, s_2, ..., s_K</td>
</tr>
<tr>
<td><strong>Parameters:</strong></td>
<td>ntree = 100, mtry = 2 (ntree: number of trees in the forest, mtry: number of variables randomly sampled as candidates at each tree split).</td>
</tr>
<tr>
<td></td>
<td>Uses out-of-sample predictions</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Uses the R package <code>glm</code>. <strong>Predictors:</strong> s_1, s_2, ..., s_K</td>
</tr>
<tr>
<td><strong>Parameters:</strong></td>
<td>kernel = “radial”, method = “C-classification” (for classification vs regression). The remaining parameters were set to their default values.</td>
</tr>
<tr>
<td>Support vector machines</td>
<td>Uses the R package <code>e1071</code>. <strong>Predictors:</strong> s_1, s_2, ..., s_K</td>
</tr>
<tr>
<td><strong>Parameters:</strong></td>
<td></td>
</tr>
<tr>
<td>Naive Bayes smoothing</td>
<td>Model: P(GT</td>
</tr>
<tr>
<td></td>
<td>( s_k = \frac{1}{\sqrt{</td>
</tr>
<tr>
<td></td>
<td>( \epsilon^2 ) is estimated using the DerSimonian and Laird method [52]</td>
</tr>
<tr>
<td></td>
<td>Naive Bayes. ( P(GT) ) was set to the value 0.5</td>
</tr>
<tr>
<td>Arithmetic average</td>
<td>Model: ( \tau = \frac{s_1 + s_2 + ... + s_K}{K} )</td>
</tr>
<tr>
<td>Geometric average</td>
<td>Model: ( \tau = \sqrt{s_1 s_2 ... s_K} )</td>
</tr>
<tr>
<td>Fixed effects</td>
<td>Model: ( \tau = \frac{\bar{V} \log(\bar{V})}{\sum \bar{V}_k} )</td>
</tr>
<tr>
<td></td>
<td>( \bar{V}_k = \frac{1}{\text{VAR}(\log(V_k))} \sum \bar{V}_k )</td>
</tr>
<tr>
<td>Random effects</td>
<td>Model: ( \tau = \sum \log(V_k) / \sum V_k )</td>
</tr>
<tr>
<td>Empirical Bayes</td>
<td>Model: ( \tau = \sum \log(V_k) / \sum V_k + (1-\alpha)\beta )</td>
</tr>
<tr>
<td></td>
<td>( \alpha ) and ( \beta ) are estimated via the EM algorithm</td>
</tr>
</tbody>
</table>

s_k: signal statistic (ratio of Eq. (1)) produced from data source \( k \) for a given association (test case). s_1, s_2, ..., s_K: set signal statistics produced from data sources 1 to K for a given association. GT: ground truth assigned to a test case (true/false).

2.5. Retrospective performance evaluation

Retrospective performance evaluations were performed by using all data from a given data source up to the end of 2013 to generate signal statistics for each OMOP test case. The area under the receiver operating characteristic curve (AUC), was then used to measure the accuracy of signal detection.

Given that a large proportion of OMOP test cases were directly compiled from the scientific literature via literature review, using the OMOP benchmark to evaluate signals generated from literature derivatives is circular and would lead to fallacious results. Consequently, the MEDLINE modality was not used our retrospective evaluation of multimodal signal detection. The data sources whose signal statistics were combined in our retrospective evaluation include FAERS, our Claims data, and the Web search logs.

2.6. Backdated prospective performance evaluation

In our backdated evaluations we use all data available from a given data source up to a pre-specified time point that is prior to an ADR’s index date (Fig. 2). We define lead time as the time interval between signal detection and a ADR’s index date, and use the date on which the underlying ADR was added to the label as its index date.

We propose two performance indices called longitudinal-AUC (LAUC) and mean lead time to detection (MLT2D) that blend both accuracy and lead time to detection.

LAUC enables the assessment of signal detection accuracy as a function of the ‘earliness’ of detection, permitting statements of the form: ‘with a lead time of 36 months method A detects signals more accurately than method B’. Calculating LAUC requires backdating the analysis of each positive test case to a different time period (depending on its index date), comparing its signal score to those of each negative test case, and computing a test statistic that longitudinally generalizes to the standard AUC statistic. This is not the same as repeated AUC calculations at different time periods, unless all test cases have the same index date. Formally, for a given lead time (lt) LAUC is defined as

\[
LAUC(lt) = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \psi(s_i^+(t_i^+ - lt), s_j^-(t_j^- - lt))
\]

where

\[
\psi(x,y) = \begin{cases} 
1 & x > y \\
0.5x = y \\
0 & x < y 
\end{cases}
\]

and \( s_i^+(t), s_j^-(t) \) denote the signal score of the ith positive test case and the jth negative test case, respectively.

Fig. 2. Retrospective versus prospective performance evaluation. In the retrospective evaluation all data up to the end of 2013 is used to generate signals for historical, known associations (and negative controls) from the OMOP benchmark. A prospective evaluation is simulated by backdating the analysis to a time period preceding the recognition of a given labeled ADR by an amount specified as lead time (e.g., 36 months). At this time point signal statistics for labeled ARDs and negative controls are generated from each data source.
jth negative test case at time $t$ respectively, and $t^0_j$ the index date (e.g., labeling date) of $j$th positive test case.

MLT2D is used quantitatively how early can ADRs be identified given a pre-specified tolerance for false alert rates. For a given ADR (positive test case), lead time to detection for the ADR is the earliest time point for which the proportion of negative test cases having a signal score larger than the investigatory ADR is equal to or less than the pre-specified false alert rate. Mean lead time to detection is the average lead time to detection across all positive test cases. Building on the notation above, for a given false alert rate $\epsilon$, and a time series of lead times $l_t, \ldots, l_{t_0}$, MLT2D is formally defined as

$$MLT2D(\epsilon) = \sum_{j=1}^J l_{t_{0}^{(j)}}(\epsilon)$$

where $l_{t_{0}^{(j)}}(\epsilon)$ is the largest $l_{t}$ that satisfies

$$\frac{1}{J} \sum_{j=1}^J \psi(s^*_t(l^{0}_t-l_t),s^*_t(l^{0}_t-l_{t_0})) \geq 1-\epsilon$$

(3)

Analysis of data from search logs was excluded our backdated evaluations due to the limited time span of the data (18 months)—a result of the terms of use under which the data were collected from Web searchers.

2.7. Quantifying performance gains

To quantify the improvement gained using composite signal scores in lieu of individual scores we use both absolute and relative differences in the AUC (or LAUC) that result from the use of two competing methods. We capture relative differences using the AUC (or LAUC) proportion of error reduction defined as

$$AUC(f(S_1,\ldots,S_K)) - \max(AUC(S_1),\ldots,AUC(S_K))$$

$$1 - \max(AUC(S_1),\ldots,AUC(S_K))$$

where $K$ is the number of data sources being used for signal detection, $S_k$ is the set of signal scores generated from data source $k$, and $f(S_1,\ldots,S_K)$ is the corresponding set of composite signal scores.

The AUC proportion of error reduction relates to a family of statistical measures known as Proportional Reduction Of Error (PRE), which are based on comparing predictive error levels that result from the use of two methods of prediction [53]. The AUC proportion of error reduction has also been used in prior studies of signal detection [25,31,37].

The AUC proportion of error reduction provides an added dimension in understanding performance gains that cannot be directly observed from absolute differences alone. Arguably, an absolute AUC difference of 0.05 between one method compared to a baseline method with an AUC of 0.8 is more interesting or more significant than the same 0.05 difference compared to a baseline method with an AUC of 0.5. Arguably, it is also harder to improve on a better performing method than on a weaker performing method. The former example results in an AUC proportion of error reduction of 25% while the latter results in an AUC proportion of error reduction of 10%, thus highlighting the significance of the former improvement. Notwithstanding, because the AUC proportion of error reduction is relative to prediction error rather than prediction accuracy, it may appear to amplifies gains and should therefore be interpreted correctly. Consequently, we report and quantify performance gains in terms of both the absolute AUC difference and the proportion of error reduction.

2.8. Cross validation and significance tests

The performance of the predictive approaches for combining signals was determined using cross-validation. AUC/LAUC was determined using the recommended leave-pair-out cross-validation procedure [54], whereas MLT2D was determined using leave-one-out cross-validation.

For the former the left-out pair consists of one positive and one negative test case. For the latter, the left-out observation consists of a positive test case. In leave-pair-out cross-validation we iterate through all combinations of pairs made of one positive and one negative test case. In each iteration, the training set consists of all test cases with the exclusion of a single pair. The trained model is then applied to predict a pair of signal scores for the held-out pair, which are then input to Eq. (2) to compute AUC/LAUC.

DeLong’s test for the difference between correlated AUCs [55] (and by extension error reduction) was used to obtain one-sided $p$-values for the difference between the AUC of the multimodal (composite) signal statistics and those of the single modalities (differences significantly greater than zero). A trivial modification of DeLong’s original formulation was used to calculate cross-validated and LAUC-based $p$-values for evaluations that utilized the two. One-sided $p$-values for differences in MLT2D were obtained using paired T-tests.

3. Results

The full set of performance statistics generated for this study are provided in the Appendix. Here we summarize key findings and trends.

3.1. Retrospective performance evaluation

Fig. 3 shows a performance comparison of signal detection modalities based on the OMOP benchmark. The best performing predictive approach to combine signals resulted in an AUC of 0.87, improving the AUC of the best performing single modality (FAERS) by 0.09, which represents an AUC error reduction of 41%. The best performing weighted averaging approach resulted in an AUC improvement of 0.07, which represents an AUC error reduction of 30%. All methods to combine signals with the exclusion of two weighted averaging approaches resulted in statistically significant ($p < .05$) error reductions. Appendix A provides the full set performance statistics for this evaluation.

3.2. Prospective performance evaluation

Fig. 4 displays a performance comparison of signal detection modalities based on LAUC. The evaluation was performed for lead times ranging from 12 months to 72 months in 6 month intervals. The figure shows that the class-leading methods for combining signals result in LAUCs in the range 0.78–0.86, AUC improvements in the range 0.04–0.09, which represent error reductions ranging from 17% to 37%. For example, given a lead time of 48 months, the multimodal approach provides an AUC improvement of 0.09 over MEDLINE and FAERS,
which represents an AUC error reduction of 37%. For the majority of
lead times, the LAUC error reductions provided by these leading
methods are statistically signiﬁcant. Appendix B provides the full set of
performance statistics for this LAUC evaluation.

Fig. 5 shows a comparison of MLT2D between modalities as a func-
tion of eight representative false alert rate cutoffs ranging from 1% to
30%. A total of 24 time points ranging from a lead time of 0 to 81
months, with three month intervals, were used to form a time series
of signal scores for each test case, and a three-point moving average was
used to smooth short-term signal score ﬂuctuations.

Expectedly, the ﬁgure shows that lead time to detection increases as
our tolerance for false alerts increases. For example, with a false alert
rate of 5% the multimodal approach detects ADRs 3.5 years in advance
of our index dates (labeling revision dates). A false alert tolerance of
30% would add one more year of lead time. Relative to our index dates,
the added lead time provided by the multimodal approach over the
MEDLINE modality ranges from 7 to 10 months, and 7 to 22 months
over the FAERS modality. The class-leading methodologies to combine
signals provided virtually equivalent performance, with statistically
signiﬁcant improvements over the single modalities for the majority of
false alert rates.

The distribution of lead times resulting in MLT2D for the two class-
leading methodologies is illustrated in Fig. 6. The ﬁgure shows that the
proportion of test cases with lead times at the time series boundaries
(0–81 months) decreases from about 50% to 30% as we vary the false
alert tolerance from 1% to 30%. The ﬁgure also shows that about 10% of
our test cases are undetectable (lead time = 0) and about 10% are always
detectable (lead time = 81) regardless of the false alert tolerance.

Fig. 5. Mean lead time to detection as a function of eight representative false alert rates. Combined-PR: combined signals produced by the best performing predictive algorithm. Combined-WA: combined signals produced by the best performing weighted averaging approach. Note that the scale of the ﬁrst two points on the x-axis is different from the remaining points.

Fig. 6. Cumulative distribution of lead times to detection for the class-leading methods to combine signals, given false alert rates of 1% and 30%. The lower the curve, the more probability mass attached to greater lead times. FAR: false alert rate. Combined-PR: combined signals produced by the best performing predictive algorithm. Combined-WA: combined signals produced by the best performing weighted averaging approach.

Appendix B provides the full set of performance statistics for the
MLT2D evaluation, and an equivalent evaluation for median lead time
to detection, with similar results, is also provided in Appendix B.

3.3. Method differences

The predictive class of approaches to combine signals generally
performed better than the weighted averaging class of methods, but the
performance differences between the class leaders were small and sta-
tistically insigniﬁcant. The within class performance variation was
larger for the weighted averaging class, and while there was no clear
winner among methods within the predictive class, their performance
differences were minimal. The arithmetic and geometric average ap-
proaches resulted in near equivalent performance, usually better than
their other class members. This was because the weights attached to
each source, which are proportional to the amount of data available for
a given association, was inversely related to predictive power of the
data sources. Table 3 displays the best performing methods for each
evaluation category. With the exclusion of FAERS and search logs in the
retrospective evaluation, the point estimate $\hat{s}$ (Eq. (1)) provided greater
performance than its 5% lower limit.
### Table 3

<table>
<thead>
<tr>
<th>Combined-PR</th>
<th>RF</th>
<th>LR</th>
<th>SVM</th>
<th>Combined-PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td></td>
<td></td>
<td></td>
<td>LAUC</td>
</tr>
<tr>
<td>Retrospective</td>
<td></td>
<td></td>
<td></td>
<td>MILT2D</td>
</tr>
</tbody>
</table>


4.3. Methods

To demonstrate the potential of multimodal signal detection we chose to highlight peak performance, and we acknowledge that our reported performance statistics should be critically judged as such. Nevertheless, the anticipated improvements of multimodal signal detection were consistently seen across a large number of methodologies we evaluated (with small differences, see Appendix), and across two distinctive benchmarks. It also clear that the methods we selected are only a sample of possible methods to generate and combine signals, and others should also be considered.

Although we found that the class of predictive approaches to combine signals generally resulted in greater performance and were more consistent, the approaches based on weighted averaging provide stronger support for the paradigm given that they are not fitted to the evaluation benchmarks. Their relative simplicity, ease of implementation, and somewhat surprisingly good performance, makes them attractive candidates for future implementations.
5. Conclusions

Within the scope of our evaluation framework, the results support the notion that utilizing and jointly analyzing multiple data sources may lead to improved signal detection. However, given several limitations associated with the data and benchmarks used in this work, it is currently not possible to make definitive statements about the ultimate utility, or quantifiable benefits of, or the implementation of multimodal signal detection. By highlighting the challenges and limitations involved in the construction and evaluation of multimodal signal detection, by exploring methodologies to combine signal statistics, and by developing methods to evaluate multimodal signal detection, the work embodied in this manuscript brings us closer to understanding the promise of multimodal signal detection. Further development requires deeper understanding of the data sources used in the construction of a multimodal system, their inherent biases and limitations, their relationships and interactions, and the mutual properties that would make them perform well in combination. The applicability of additional data sources must also be examined, as well as the utility of other methods to generate and synthesize signal statistics. Crucially, additional benchmarks will need to be developed to expand the scope of evaluated surveillance scenarios, to quantify gains from multimodal approaches, and to further characterize its potential benefits. Given the complexity of ADRs, benchmarks can only provide close approximations to the true performance characteristics of signal detection methodologies. To this end, our performance statistics should be interpreted as such approximations. Finally, our work contributes new performance indices that blend both the accuracy and the lead-time-to-detection. These indices are general enough to support other studies of signal detection.

Disclaimer

The findings and conclusions expressed in this report are those of the authors and do not necessarily represent the views of the U.S. FDA or the federal government.

Acknowledgments

This work was partially supported by the National Institute of General Medical Sciences (NIGMS) grant R01 GM101430, and the intramural Research Program of the NIH, National Library of Medicine (NLM). This work also received support from the U.S. Food and Drug Administration (FDA) through the Center for Drug Evaluation and Research (CDER) Critical Path Program [interagency agreement with NLM (XLM12011 001)], from the Office of Translational Sciences at CDER, from the UCSF-Stanford CERSI award (U01FD004979), as well as from the ORISE Fellowship program at the FDA.

Competing financial interests

The authors declare no competing financial interests. Rave Harpaz and William DuMouchel are employed by Oracle Health Sciences. Nigam H Shah is a Science Advisor to Apixio Inc. and Learning Health Inc. Carol Friedman consults for Health Fidelity Inc. Martijn Schuemie is employed by Janssen Research and Development. Eric Horvitz and Ryn W White are employed by Microsoft.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.10.013.


