Journal of the American Medical Informatics Association, 24(5), 2017, 913–920
doi: 10.1093/jamia/ocx022
Advance Access Publication Date: 21 March 2017
Research and Applications

Research and Applications

Development of an automated assessment tool for MedWatch reports in the FDA adverse event reporting system

Lichy Han,[1] Robert Ball,[2] Carol A Pamer,[2] Russ B Altman,[3,4] and Scott Proestel[2]

[1]Biomedical Informatics Training Program, Stanford University, Stanford, CA, USA, [2]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA, [3]Department of Genetics, Stanford University and [4]Department of Bioengineering, Stanford University

Corresponding Author: Scott Proestel, Division of Epidemiology, Office of Biostatistics and Epidemiology, FDA Center for Biologics Evaluation and Research, White Oak Building 71, Room 1260, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA. Phone: (240) 402-0396. E-mail: Scott.Proestel@fda.hhs.gov

Received 1 September 2016; Revised 30 January 2017; Accepted 24 February 2017

ABSTRACT

Objective: As the US Food and Drug Administration (FDA) receives over a million adverse event reports associated with medication use every year, a system is needed to aid FDA safety evaluators in identifying reports most likely to demonstrate causal relationships to the suspect medications. We combined text mining with machine learning to construct and evaluate such a system to identify medication-related adverse event reports.
Methods: FDA safety evaluators assessed 326 reports for medication-related causality. We engineered features from these reports and constructed random forest, L1 regularized logistic regression, and support vector machine models. We evaluated model accuracy and further assessed utility by generating report rankings that represented a prioritized report review process.
Results: Our random forest model showed the best performance in report ranking and accuracy, with an area under the receiver operating characteristic curve of 0.66. The generated report ordering assigns reports with a higher probability of medication-related causality a higher rank and is significantly correlated to a perfect report ordering, with a Kendall's tau of 0.24 ($P = .002$).
Conclusion: Our models produced prioritized report orderings that enable FDA safety evaluators to focus on reports that are more likely to contain valuable medication-related adverse event information. Applying our models to all FDA adverse event reports has the potential to streamline the manual review process and greatly reduce reviewer workload.

Key words: drug-related side effects and adverse reactions, supervised machine learning

BACKGROUND AND SIGNIFICANCE

The US Food and Drug Administration (FDA) receives more than 4000 medication safety reports every day, and the number of reports received each year has been increasing exponentially over the last decade. These reports are stored in a database known as the FDA Adverse Event Reporting System (FAERS), which has collected over 11 million reports since its inception in 1969.[1] In the United States, reporting these adverse events, medication errors, and product quality issues by health care professionals and consumers via the MedWatch program is voluntary, but it is mandatory for drug manufacturers.[2] The FDA uses these reports to detect safety issues that may not have been identified during pre-market clinical trials used as the basis for medication approval. Among the reasons for not detecting safety issues during pre-market evaluation are that the

Published by Oxford University Press on behalf of the American Medical Informatics Association 2017.

adverse event may be extremely rare or may be occurring in a population of individuals who were not previously studied.[3,4]

While the volume of reports is significant and increasing yearly, many of the reports do not provide sufficient information to determine whether the suspect drug caused the reported adverse event, or the information provided does not reasonably suggest that the adverse event was caused by the suspect medication. Therefore, developing a tool that could assist FDA safety evaluators by identifying reports that are most likely to be useful would be highly valuable and could help to ensure that safety issues do not go undetected.

There has been considerable interest in the use of natural language processing (NLP) and machine learning to enable workers to focus on subject matter that is most likely to be useful.[5,6] Free-text narratives often contain details that allow causal inference, so extracting these details is expected to be an important aspect of this work. In the current project, we explore the possibility of using NLP and machine learning to assess the likelihood of drug causality for safety reports submitted to the FDA. In this first phase, we demonstrate the feasibility of and potential benefits to the adverse event review process in a subset of adjudicated FAERS reports. With success, FDA safety evaluators would be able to apply our system to the entire corpus of over 11 million reports, enabling them to focus their review efforts on reports that are most likely to indicate the emergence of a new safety concern.

METHODS

Data and gold standard

A "gold standard" for drug causality assessment was created by deidentifying and redacting 326 case reports from FAERS. Safety reports were selected from FAERS based on convenience sampling of cases received by the FDA between November 1997 and March 2015. Specifically, the redacted fields include the patient identification number; date of birth; reporter name, organization, and location; and dates in the narrative. Only cases received after November 1997 were eligible, as only those reports contain an electronically retrievable and machine readable narrative section. The Stanford University Institutional Review Board approved this study (IRB-34866). Access to the data used in this study for research purposes can be requested through the FDA Technology Transfer Program at techtransfer@fda.hhs.gov.

Every case report was assessed for causality by 3 FDA safety evaluators using a modified version of the World Health Organization–Uppsala Monitoring Centre (WHO-UMC) criteria for drug causality assessment (Table 1). All evaluators participated in training sessions to ensure consistent scoring. Evaluators were instructed to consider the totality of the case report in their assessment (ie, both structured and unstructured data). If 2 or more evaluators agreed on the assessment, then the assessment of the majority was used. If all 3 evaluators disagreed, a final adjudication was made by a fourth evaluator. An assessment of causality was based on only the suspect medication(s) and adverse reaction(s) identified by the reporter of the event. Other products or reactions included in the report, but not identified as "suspect" by the reporter, were not assessed. As multiple medications and adverse reactions could appear in a single report, the case assessment was based on the medication-event combination that gave the highest likelihood of causality assessment based on the WHO-UMC scale.

After adjudication, case reports were randomly divided by the FDA investigators into a training and a test set, consisting of 60% and 40% of the data, respectively. We chose these proportions in order to capture a more comprehensive and representative set of the causality categories in the test set for evaluation. We then formulated our problem as a binary classification task by aggregating the causality categories into 2 groups: (1) *Certain*, *Probable*, *Possible* and (2) *Unlikely, Unassessable*. All features and models were built using the training data and assessed using the test data. All analyses were performed using R 3.3.0 (R Development Core Team, Vienna, Austria).

Feature engineering

Adverse event reports consist of structured data and an unstructured narrative. The structured data consists of fields such as age, adverse event outcomes (death, hospitalization, other), timeline of report entry by the FDA, adverse event reporters and their qualifications, terms mapped to the Medical Dictionary for Regulatory Activities (MedDRA®), and drug suspect products. The unstructured narrative is a free-text entry that varies in length from one word or sentence

Table 1. Modified WHO-UMC causality categories

Causality Term	Assessment Criteria
Certain	Event or laboratory test abnormality, with plausible time relationship to drug intake Cannot be explained by disease or other drugs Response to withdrawal plausible (pharmacologically, pathologically) Event definitive pharmacologically or phenomenologically (ie, an objective and specific medical disorder or a recognized pharmacologic phenomenon) Rechallenge satisfactory, if necessary
Probable/likely	Event or laboratory test abnormality, with reasonable time relationship to drug intake Unlikely to be attributed to disease or other drugs Response to withdrawal clinically reasonable Rechallenge not required
Possible	Event or laboratory test abnormality, with reasonable time relationship to drug intake Could also be explained by disease or other drugs Information on drug withdrawal may be lacking or unclear
Unlikely	Event or laboratory test abnormality, with a time to drug intake that makes a relationship improbable (but not impossible) Disease or other drugs provide plausible explanation
Unassessable[a]	Cannot be judged because information is insufficient or contradictory
Medication Error[a]	Report suggesting accidental or intentional inappropriate use Not necessarily associated with an adverse event
Product Quality Issue[a]	Report suggesting a possible product quality issue Not necessarily associated with an adverse event

[a] Category modified from the WHO-UMC Causality Assessment Scale. The unmodified version of this scale is located at http://who-umc.org/Graphics/26649.pdf.[7]

to multiple paragraphs. These narratives typically detail the chronological course of the adverse event, occasionally including medications and laboratory results. Protected health information was redacted from both the structured data and unstructured narratives prior to review by the Stanford investigators.

Structured data

We processed the structured data by re-encoding categorical variables, generating new features, and summarizing existing data. We expanded each categorical variable by transforming the possible values into multiple binary features. Age was binned by decade, and medications were binned by the first level of the Anatomical Therapeutic Chemical (ATC) Classification System.[8] For the report types, direct reports are sent to the FDA by the general public, whereas 15-day (expedited) and nonexpedited reports are sent by pharmaceutical companies. The expedited 15-day time frame is for adverse events that are serious and unexpected. We calculated the number of days to complete the report as the data entry completion date minus the initial FDA received date. The data entry completion date is updated whenever a follow-up report is added to the case report. The total number of outcomes and reporters was calculated and included as additional features. MedDRA terms were represented by the number of preferred terms (PTs), high-level terms, high-level group terms, and system organ classes associated with each report.[9]

Unstructured narrative

The unstructured narratives were tokenized, lemmatized, and part-of-speech tagged using Stanford CoreNLP.[10] Tokenization refers to splitting the narrative into individual words, lemmatization refers to converting each word into its base form, and part-of-speech tagging refers to assigning a part of speech (eg, noun, verb) to each word. We incorporated the number of sentences and average number of words, nouns, verbs, adjectives, and adverbs per sentence as nonsemantic features. We additionally counted the number of redacted terms in each narrative, which typically corresponded to dates detailing the time course of the adverse event.

Expert opinion derived features

From the structured data, we compiled all MedDRA preferred terms that were present in at least 5 reports. We surveyed 5 adjudicators, who assessed each of these preferred terms and indicated if the presence of the term in a report would tend to increase the likelihood of a medication-related adverse event. Any preferred term that had at least 3 affirmative votes was included as an additional binary variable. We refer to this binary feature as the "presence of curated PTs." The included terms were "drug interaction," "acute kidney injury," "seizure," "drug hypersensitivity," "drug ineffective," "rhabdomyolysis," "product quality issue," and "toxicity to various agents."

From the unstructured narratives, 5 adjudicators compiled a list of 10 words and phrases that they believed were more likely to be associated with a medication-related adverse event. These words and phrases were "toxicity," "induced," "rechallenge," "probable," "no alternative etiology," "reaction," "adverse reaction," and "drug level." Alternative spellings of these chosen phrases were included as well. The presence of any of these words and phrases in the narrative was included as a binary feature, titled "presence of curated terms."

Model construction and evaluation

We constructed models to differentiate reports with an assessment of *Certain*, *Probable,* or *Possible* from reports with an assessment of *Unlikely* or *Unassessable*. We built models using L1 regularized logistic regression, random forest, and support vector machines (SVMs) using the *glmnet*,[11] *randomForest*,[12] and *caret*[13] packages in R, respectively. Model parameters were chosen by 10-fold cross-validation[14,15] on the training set; no reports from the test set were used in model parameter selection or model construction. The penalty parameter lambda for the L1 regularized logistic regression model was chosen as the largest value of lambda that was within 1 standard error of the minimum error.[15] The values of lambda searched ranged from 0.0001 to 0.15, as determined by the *glmnet* package. A grid search was used to determine the optimal values of C and gamma for our SVM model. We searched over 10 values for each parameter, with C ranging from 0.01 to 5 and gamma ranging from 0.001 to 2. Our model predictions were evaluated by calculating the sensitivity, positive predictive value, accuracy, and area under the receiver operating characteristic curve (AUROC). To assess the relative importance of our features, we extracted features retained in the L1 regularized logistic regression model and used the mean decrease in the Gini index in the random forest model.

We further evaluated the utility of our models by using their predictions to create ranked lists of reports corresponding to the order in which they would be reviewed by manual adjudication. Specifically, for all 3 classifiers, we ordered the reports by their probability of having an assessment from *Certain* to *Possible*, where a greater probability would correspond to a report being adjudicated sooner and having an earlier ranking. For comparison, we made a ranked list of reports based on the date of data entry completion by the FDA, which we used to represent manual review of reports on a first come, first reviewed basis. We also assessed ranking the reports in reverse date order to simulate a scenario where the newest reports would be reviewed first. Additionally, we simulated arbitrary orderings by generating 10 000 random report rankings. We compared each of these report rankings to a perfectly ranked list using Spearman's rho correlation coefficient, Kendall's tau correlation test, and average precision (AP). We calculated AP using 2 cutoffs: (1) the entire test set ($N = 125$) and (2) the number of reports with an assessment of *Certain*, *Probable*, or *Possible* ($N = 64$).

We performed an error analysis of our model results by categorizing the false positive (FP) and false negative (FN) errors based on our features. We considered reports with an assessment of *Certain, Probable,* or *Possible* as positives and those with an assessment of *Unlikely* or *Unassessable* as negatives. We compared the distribution of each feature for the accurately classified reports to the reports classified as FN or FP. For categorical variables, we used the chi-squared test, and for continuous variables, we used Student *t*-test.

RESULTS

Data

Of the 326 reports, approximately half had an assessment of at least *Possible* (Figure 1), and no reports in the test set had an assessment of *Certain.* There was a total of 25 reports with an assessment of *Medication Error* or *Product Quality Issue*, 6 from the test set and 19 from the training set. Focusing on reports with an assessment of *Certain* to *Unassessable* resulted in a total of 301 reports, for a training set of 176 (107 *Certain, Probable, Possible*; 69 *Unlikely, Unassessable*) and test set of 125 (64 *Certain, Probable, Possible*; 61 *Unlikely, Unassessable*).

When assessing adjudicator agreement, we found that at least 2 of the 3 adjudicators tended to agree on any given report. At the granular level, using 7 classes, all 3 adjudicators agreed on 37.4% of

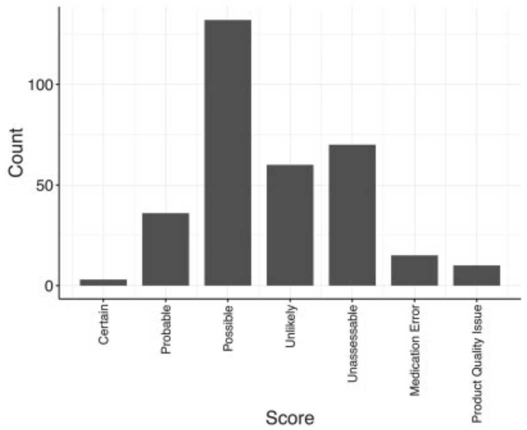

Figure 1. Histogram of FAERS report assessments.

the reports, and 92.6% of reports had at least 2 adjudicators in agreement. When we grouped the reports (*Certain, Probable, Possible*, vs *Unlikely, Unassessable*), all 3 adjudicators agreed on 61% of the reports.

Features

Summary statistics for features constructed from structured and unstructured data fields for reports with assessments between *Certain* and *Unassessable* are shown in Table 2. Features from structured data were constructed in separate categories, including report logistics, reporters, MedDRA terms, outcomes, and drugs. The unstructured narratives had an average length of 366.1 ± 444.7 words, with a range from 1 to 2575. We found that the "Days to Complete Report" feature correlated with the number of follow-up reports (data not shown) and potentially more information about the case. We chose to use "Days to Complete Report" to represent the length of the case timeline due to the increased granularity of dates over the number of reports. Overall, the training and test sets are similar across our features, though we did observe minor differences in the distribution of drug ATC classes and the number of days to complete the report.

Model evaluation

Results from our models on the test set of reports are shown in Table 3. The parameters chosen by cross-validation for our models were lambda = 0.0646, C = 1, and gamma = 0.05. The random forest and L1 regularized logistic regression models attained the highest AUROC (Figure 2), with the random forest model resulting in slightly higher accuracy. The random forest model also showed the greatest similarity to the perfect ranking and the most improvement over other report orderings, with a Kendall's tau of 0.24 ($P = .002$) and Spearman's ρ of 0.28. Perfect ranking resulted in AP values of 0.85 ($N = 125$ cutoff) and 1.00 ($N = 64$ cutoff), and out of all our models and heuristics, the random forest model attained the closest performance, with AP values of 0.65 and 0.74, respectively. Neither of the date-ranking heuristics was significantly correlated with the perfect ranking. Reverse ranking achieved approximately the inverse of the results of the forward ranking, the difference due to multiple reports with the same assessment in the test set. Although the SVM and L1 regularized logistic regression models showed similar accuracy and AUROC, the report rankings they produced showed no significant correlation to the perfect ordering.

Using the report rankings from the random forest model resulted in a shift of reports with lower assessments being ranked earlier, whereas ordering reports by date resulted in these reports being more concentrated at the end of the ordering (Figures 3A and B). As expected, the random orderings tended to produce uniform assessment distributions, with correlation values around 0 (Figures 4A and B). Our random forest model produced a more significantly correlated report ordering than all but 20 of the 10 000 random ordering simulations and outperformed both date heuristics (Figures 4A and B). Using our random forest model, in order to review at least 80% of the reports with an assessment of *Certain, Probable,* or *Possible* in the test set, evaluators would need to review 86 out of the 125 reports, as compared to 100 out of 125 if we reviewed by date or an average of 99 out of 125 if we reviewed in random order. This represents a potential 10% reduction in workload at a sensitivity threshold of 80%. Further examination at the individual assessment level showed a shift of reports with an assessment of *Probable* or *Possible* in the test set to earlier in the report order, and reports with an assessment of *Unlikely* or *Unassessable* to later (Figure 5). Notably, the magnitude of the change in report order appeared to be more accentuated for reports with a more extreme assessment of *Probable* or *Unassessable.*

When examining the FP errors, we found that there were no statistically significant differences among the features. Out of all the features, the largest difference between FP errors and accurately classified reports was in the number of direct type reports. We found that none of the FP errors were direct reports, whereas in the training set, a direct report type tended to be associated with the positive class.

For the FN errors, we found that these reports had significantly less manually curated PT terms from MedDRA ($P = .007$) and were significantly more likely to have death as an outcome ($P = .03$). None of the FN reports, despite being evaluated as having a higher likelihood of medication-related causality, contained any of the PT terms that were manually curated. Additionally, the FN errors consisted of 2 out of a total 16 *Probable* reports and 11 out of 48 *Possible* reports. Thus, the classifier tended to misclassify reports with a lower causality assessment as negative.

Feature importance

We extracted important features from our random forest and L1 regularized logistic regression models (Table 4). The latter model included 6 variables and the former model included correlated variables and features from both the structured data and the unstructured narratives.

DISCUSSION

A key component of FDA regulatory activities is pharmacovigilance, which partly relies on post-marketing surveillance and spontaneous reporting systems, including the FDA Adverse Event Reporting System. Over the last decade, the number of adverse event reports has increased exponentially, resulting in a substantial workload for reviewers. Delays in detecting drug adverse events can have costly and detrimental effects on public health, and thus a system to identify reports most likely to contain information demonstrating causal drug events would be highly beneficial. Researchers have investigated such approaches using the US Vaccine Adverse Event Reporting System, in which extracted text features[16,17] were used with multiple classification algorithms to create an effective report classification model.[18–20]

The success of text classification in the Vaccine Adverse Event Reporting System and previous computational discoveries of new medication-related adverse events in FAERS[21–27] have generated significant interest in developing a classification system for FAERS. To accomplish this, we built models to classify and rank adverse event

Table 2. Summary of features derived from structured and unstructured data

		Training Set	Test Set
Structured Data	No. of Reports	176	125
	Age	50.5 ± 23.0	54.6 ± 17.9
	Sex: Male, n (%)	75 (42.6)	55 (44.0)
	Days to Complete Report		
	Report type	91.5 ± 305.6	55.9 ± 92.4
	15-day, n (%)	134 (76.1)	86 (68.9)
	Non-expedited, n (%)	32 (18.2)	28 (22.4)
	Direct, n (%)	10 (5.7)	11 (8.8)
	Reporters		
	Health professional, n (%)	37 (21.0)	22 (17.6)
	Consumer, n (%)	10 (5.7)	10 (8.0)
	Foreign, n (%)	20 (11.4)	8 (6.4)
	Other, n (%)	120 (68.2)	85 (68.0)
	MedDRA Terms		
	# PTs	3.9 ± 5.2	3.1 ± 2.7
	# HLTs	3.2 ± 2.7	2.9 ± 2.4
	# HLGTs	2.9 ± 2.4	2.8 ± 2.1
	# SOCs	2.5 ± 1.7	2.3 ± 1.7
	Presence of curated PTs, n (%)	18 (10.2)	12 (9.6)
	Outcomes		
	Death, n (%)	17 (9.7)	12 (9.6)
	Hospitalization, n(%)	71 (40.3)	40 (32.0)
	Other, n (%)	76 (43.2)	62 (49.6)
	Serious outcome, n (%)	140 (79.5)	98 (78.4)
	Number of Drug Suspects		
	Drug suspect ATC class, first level	2.01 ± 1.9	2.06 ± 2.4
	J (antiinfectives for systemic use), n (%)	30 (17.0)	9 (7.2)
	R (respiratory system), n (%)	12 (6.8)	10 (8.0)
	A (alimentary tract and metabolism), n (%)	43 (24.4)	33 (26.4)
	N (nervous system), n (%)	29 (16.4)	30 (24.0)
	C (cardiovascular system), n (%)	10 (5.7)	12 (9.6)
	L (antineoplastic and immunomodulating agents), n (%)	31 (17.6)	19 (15.2)
Unstructured Narrative	No. of Sentences		
	# Words per sentence	23.1 ± 24.2	19.6 ± 22.3
	Parts of speech	18.1 ± 7.4	17.8 ± 10.1
	# Nouns per sentence	6.6 ± 3.2	6.3 ± 4.2
	# Verbs per sentence	2.6 ± 1.0	2.7 ± 1.4
	# Adjectives per sentence	1.8 ± 1.0	1.7 ± 1.3
	# Adverbs per sentence	0.5 ± 0.3	0.5 ± ± 0.4
	Number redacted	1.2 ± 3.0	0.7 ± 1.9
	Presence of curated terms	50 (28.4)	36 (28.8)

Binary features are reported using the number and percentage of reports, and numerical features are reported using mean ± standard deviation.

Table 3. Performance metrics of classification models

Metric	By Date	Reverse Date	Random Order ($N = 10{,}000$)	Random Forest	L1 Regularized Logistic Regression	SVM
Sensitivity	–	–	–	0.66	**0.69**	0.50
PPV	–	–	–	**0.64**	0.58	0.60
Accuracy	–	–	–	**0.63**	0.58	0.58
AUROC	–	–	–	**0.66**	**0.66**	0.58
Spearman's ρ	−0.09	0.11	Mean (SD): −0.002 (0.09)	**0.28**	−0.04	0.01
Kendall's tau	−0.07	0.09	Mean (SD): −0.002 (0.08)	**0.24**	−0.04	−0.03
P-value	.31	.23	Mean (SD): 0.50 (0.27)	**0.002**	0.62	0.73
AP ($N = 125$)	0.46	0.57	Mean (SD): 0.51 (0.04)	**0.65**	0.47	0.53
AP ($N = 64$)	0.44	0.59	Mean (SD): 0.51 (0.07)	**0.74**	0.44	0.55

Random report orderings are summarized using the mean and standard deviation (SD) of the correlation metrics.
For each metric, the bolded value indicates the best performing classifier.

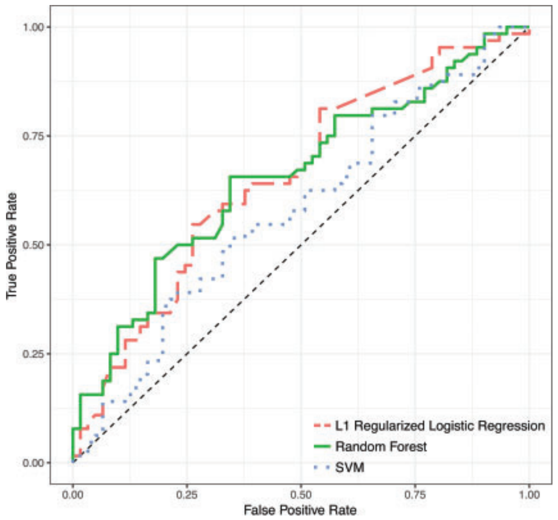

Figure 2. ROC curves for all classification models.

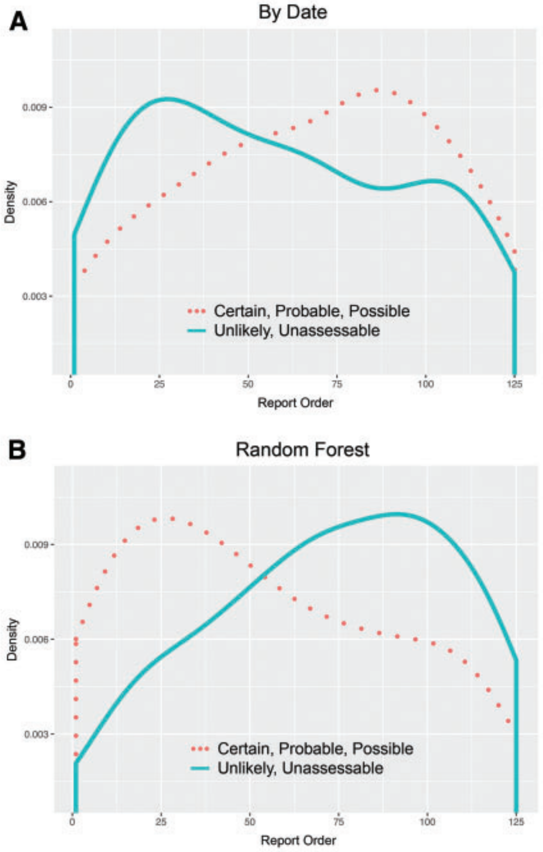

Figure 3. Comparison of report orderings in the held-out test set by (**A**) date and (**B**) random forest with assessments of *Certain, Probable,* or *Possible* vs assessments of *Unlikely* or *Unassessable.*

reports based on the likelihood of medication-related causality. In addition, we showed the potential utility of our models to assist manual adjudicators by shifting reports with a higher probability of medication-related causality to a higher priority in rank order.

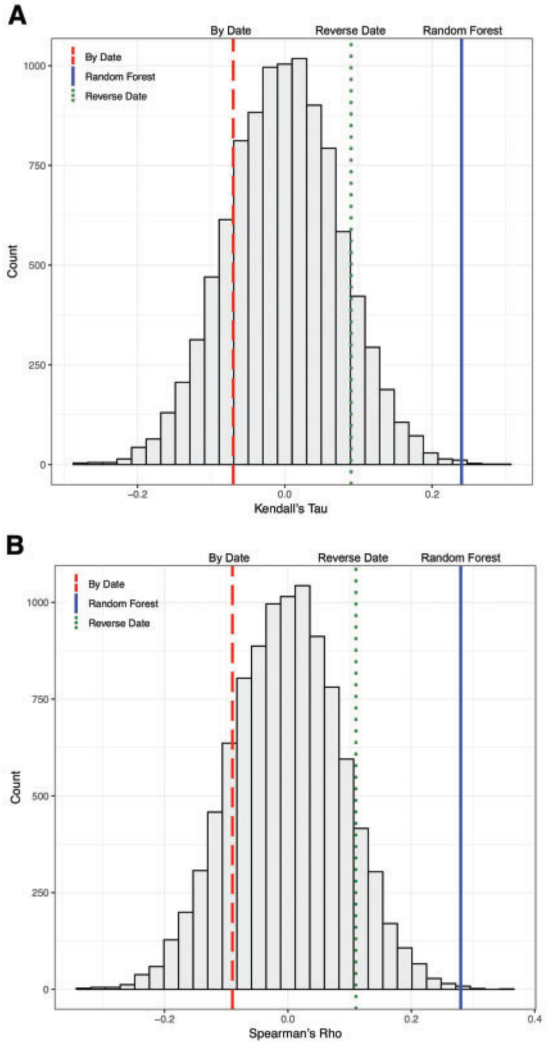

Figure 4. Histograms of (**A**) Kendall's tau and (**B**) Spearman's rho correlation coefficients for the 10 000 random report orderings.

For the first phase of this study, we chose to focus on reports with assessments of *Certain* to *Unassessable*, as they constituted over 90% of our corpus. Though the modified WHO-UMC causality categories additionally include *Medication Error* and *Product Quality Issue*, these reports were heterogeneous and few in number. Furthermore, *Medication Error* and *Product Quality Issue* are MedDRA preferred terms that appear as part of the MedDRA structured data fields. These can aid in the detection of reports with these issues, whereas a system for differentiating reports with assessments from *Certain* to *Unassessable* does not yet exist.

We engineered features based on structured data fields and extracted nonsemantic features from unstructured narratives. We chose to focus on nonsemantic features for this initial corpus of reports due to the wide variety of narratives, which range from one word to multiple paragraphs. Due to this, we found that patterns of words and dependencies were frequently unique and that such patterns would not generalize to the entirety of the adverse event report database. Instead, we relied on expert knowledge to curate terms that are considered to raise the probability of a report having medication-related causality. We recognize that the lack of semantic features is a feature engineering limitation, mostly due to the size of our current corpus. However, we believe that the unstructured narratives contain additional valuable information, and with the entirety of FAERS, we plan

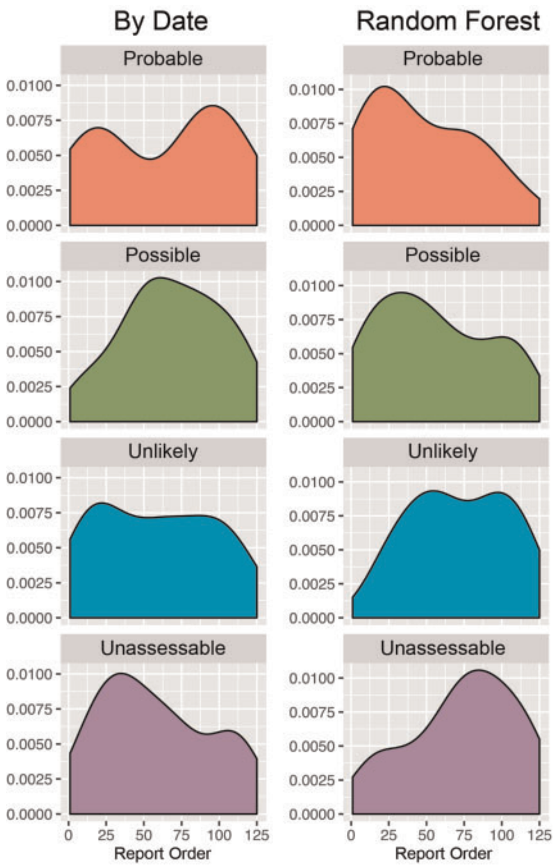

Figure 5. Individual assessment level density plots based on reports ordered by date and by using the random forest model in the held-out test set. There were no reports with an assessment of *Certain* in the test set.

to fully leverage semantic features to improve the discrimination power of our models. Examples of such features include entity recognition using the Unified Medical Language System,[28] quantification of common words and phrases using term frequency, and relationship extraction between words at the sentence level.

With our engineered features, we were able to see modest separation between high- and low-probability reports using 3 different types of machine learning models. The superior results of the L1 regularized logistic regression and random forest models over the SVM model suggest that data-driven feature selection is beneficial in avoiding overfitting and improving performance. When we assessed the reports misclassified by our model, we found that the FP and FN errors had features generally opposite to the predictive trends present in the training data. We attribute these errors to mild overfitting, which is due, in part, to the limitations imposed by a small corpus.

Upon closer inspection of the features selected by our models, we found that the L1 regularized logistic regression model retained only 6 features, many of which were also considered highly important by the random forest model. The majority of these features appear to have an intuitive interpretation, such as a consumer reporter increasing the likelihood of a report being *Unlikely* or *Unassessable*. As the "Days to Complete Report" feature reflects the addition of follow-up reports, a longer report completion time appears to increase the likelihood of a *Certain*, *Probable*, or *Possible* assessment. Interestingly, these features range across multiple feature categories, including outcomes, MedDRA terms, drug suspects, and reporter qualifications. The feature selection suggests that each of these categories likely contains correlated variables, with each category providing additional value for classification.

In addition to report classification, we further assessed utility based on reordering reports to expedite the review process for manual adjudicators. We see high-probability reports shifting to earlier in rank order at the individual assessment report level, indicating that our model can improve upon the current system of reviewing by date or randomly. We acknowledge that our work is limited by the number of reports and has the potential for additional feature extraction from the unstructured narratives. This is due to the time-intensive nature of adjudication and curation, which also serves as the motivation for this work. Despite these limitations, we believe that the benefits seen in this work can be magnified upon expansion of our model to the millions of reports in the FAERS system. Furthermore, this evaluation simulates a more realistic application, which can be easily integrated into the existing workflow. Its implementation has the potential to increase efficiency and improve detection of safety issues by prioritizing reports with a higher probability of medication-related causality.

CONCLUSION

This study serves as the foundation for construction of a system to detect adverse event reports with a high probability of indicating medication-related causality. We explored and constructed several models, which we used to demonstrate feasibility and applicability in aiding manual reviewers via report prioritization. Expansion of these models to the entirety of the FDA adverse event reporting

Table 4. Important features included by the random forest and L1 regularized logistic regression models

Random Forest				L1 Regularized Logistic Regression			
Feature	Mean Decrease in Gini Index	Certain, Probable, Possible	Unlikely, Unassessable	Feature	Coefficient	Certain, Probable, Possible	Unlikely, Unassessable
Age	6.312	✓		Presence of curated PTs	4.54e-3	✓	
No. of sentences	6.120	✓					
# words per sentence	5.707		✓	Outcome: death	−0.112		✓
Days to complete report	4.719	✓		Days to complete report	5.61e-6	✓	
MedDRA: # PTs	3.702	✓		Presence of curated terms	0.043	✓	
Number of drug suspects	3.053	✓		No. of drug suspects	0.011	✓	
Reporter: consumer	2.717		✓	Reporter: consumer	−0.327		✓

A check indicates that a higher value of the feature is associated with a higher likelihood of an assessment from the respective column.

system could substantially improve the manual review process and have potential downstream benefits for pharmacovigilance.

DISCLAIMER

The opinions or assertions presented herein are the private views and opinions of the authors and are not to be construed as conveying either an official endorsement or criticism by the US Department of Health and Human Services, the Public Health Service, or the Food and Drug Administration.

FUNDING

This work was supported by the FDA, which supports the University of California, San Francisco–Stanford Center of Excellence in Regulatory Sciences and Innovation, grant no. U01FD004979, and by the National Institutes of Health under award no. T32 GM007365, which supports the Stanford Medical Scientist Training Program. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Department of Health and Human Services, the National Institutes of Health, or the Food and Drug Administration.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

SP, RB, and CAP designed the study, acquired the data and gold standard, assisted with interpretation of the results, and critically revised the manuscript. LH analyzed the data, built the models, and drafted the manuscript. RBA critically revised the manuscript and made substantial contributions to interpreting the data and the results.

ACKNOWLEDGMENTS

The authors would like to thank reviewers from the FDA Center for Drug Evaluation and Research, Office of Surveillance and Epidemiology, Division of Pharmacovigilance and Division of Medication Errors and Prevention for assigning causality terms to the FAERS reports used in this study.

REFERENCES

1. US Food and Drug Administration. *FDA Adverse Event Reporting System (FAERS)*. 2016. http://www.fda.gov.laneproxy.stanford.edu/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/. Accessed November 8, 2016.
2. Kessler DA. Introducing MEDWatch. A new approach to reporting medication and device adverse effects and product problems. *JAMA*. 1993;269:2765.
3. Leape L. Reporting of adverse events. *N Engl J Med*. 2002;347:1633–38.
4. Lester J, Neyarapally GA, Lipowski E, *et al*. Evaluation of FDA safety-related drug label changes in 2010. *Pharmacoepidemiol Drug Saf*. 2013;22:302–05.
5. Almenoff J, Tonning JM, Gould a L, *et al*. Perspectives on the use of data mining in pharmaco-vigilance. *Drug Saf*. 2005;28:981–1007.
6. Duggirala HJ, Tonning JM, Smith E, *et al*. Use of data mining at the Food and Drug Administration. *J Am Med Inform Assoc*. 2015;23:428–34.
7. *The Use of the WHO-UMC System for Standardised Case Causality Assessment*. 2012. http://who-umc.org/Graphics/26649.pdf. Accessed November 8, 2016.
8. World Health Organization Collaborating Centre. *Introduction to Drug Utilization Research*. Geneva, Switzerland. 2003;1–48.
9. Brown EG, Wood L, Wood S. The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Saf*. 1999;20:109–17.
10. Manning CD, Bauer J, Finkel J, *et al*. The Stanford CoreNLP Natural Language Processing Toolkit. *Proc 52nd Annu Meet Assoc Comput Linguist Syst Demonstr.* 2014;55–60. http://aclweb.org/anthology/P14-5010. Accessed December 2, 2015.
11. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw.* 2010;33:1–22. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20808728&retmode=ref&cmd=prlinks. Accessed September 15, 2015.
12. Liaw A, Weiner M. Classification and Regression by randomForest. *R News*. 2002;2:18–22. papers3://publication/uuid/1A4C7E98-20AD-4197-BCDB-7C918F38F724. Accessed September 15, 2015.
13. Kuhn M. *Caret: Classification and Regression Training*. R package version 6.0-71. 2016. https://CRAN.R-project.org/package=caret. Accessed November 8, 2016.
14. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Int Jt Conf Artif Intell.* 1995;2:1137–43.
15. Hastie TJ, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. New York: Springer; 2009.
16. Botsis T, Buttolph T, Nguyen MD, *et al*. Vaccine adverse event text mining system for extracting features from vaccine safety reports. *J Am Med Inform Assoc*. 2012;19:1011–18.
17. Baer B, Nguyen M, Woo EJ, *et al*. Can natural language processing improve the efficiency of vaccine adverse event report review? *Methods Inf Med*. 2015;55:144–50.
18. Botsis T, Nguyen MD, Woo EJ, *et al*. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *J Am Med Inform Assoc*. 2011;18:631–38.
19. Botsis T, Woo EJ, Ball R. Application of information retrieval approaches to case classification in the vaccine adverse event reporting system. *Drug Saf*. 2013;36:573–82.
20. Botsis T, Woo EJ, Ball R. The contribution of the vaccine adverse event text mining system to the classification of possible Guillain-Barré syndrome reports. *Appl Clin Inform*. 2013;4:88–99.
21. Tatonetti NP, Ye PP, Daneshjou R, *et al*. Data-driven prediction of drug effects and interactions. *Sci Transl Med*. 2012;4:125ra31.
22. Hochberg A, Pearson R, O'Hara D, *et al*. Drug-versus-drug adverse event rate comparisons. *Drug-Safety*. 2009;32:137–46.
23. Harpaz R, Chase H, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics*. 2010;11:S7–8.
24. Xu R, Wang Q. Large-scale combining signals from both biomedical literature and the FDA Adverse Event Reporting System (FAERS) to improve post-marketing drug safety signal detection. *BMC Bioinformatics*. 2014;15:17.
25. Sakaeda T, Tamon A, Kadoyama K, *et al*. Data mining of the public version of the FDA adverse event reporting system. *Int J Med Sci*. 2013;10:796–803.
26. Poluzzi E, Raschi E, Moretti U, *et al*. Drug-induced torsades de pointes: data mining of the public version of the FDA Adverse Event Reporting System (AERS). *Pharmacoepidemiol Drug Saf*. 2009;18:512–18.
27. Wang L, Jiang G, Li D, *et al*. Standardizing adverse drug event reporting data. *J Biomed Semantics*. 2014;5:36.
28. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* .2004;32:D267–70.