

ORIGINAL REPORT

WILEY

Evaluating automated approaches to anaphylaxis case classification using unstructured data from the FDA Sentinel System

Robert Ball¹ | Sengwee Toh² | Jamie Nolan² | Kevin Haynes³ | Richard Forshee⁴ | Taxiarchis Botsis⁴

¹Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, FDA, Silver Spring, MD, USA

²Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA

³Translational Research for Affordability and Quality, HealthCore, Inc., Wilmington, DE, USA

⁴Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research, FDA, Silver Spring, MD, USA

Correspondence

R. Ball, Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, FDA, Silver Spring, MD, USA.
Email: robert.ball@fda.hhs.gov

Present Address

Taxiarchis Botsis, The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Funding information

US FDA, Grant/Award Number: HHSF22301012T

Abstract

Introduction: In May 2008, the Food and Drug Administration launched the Sentinel Initiative, a multi-year program for the establishment of a national electronic monitoring system for medical product safety that led, in 2016, to the launch of the full Sentinel System. Under the Mini-Sentinel pilot, several algorithms for identifying health outcomes of interest, including one for anaphylaxis, were developed and evaluated using data available from the Sentinel common data model.

Purpose: To evaluate whether features extracted from unstructured narrative data using natural language processing (NLP) could be used to classify anaphylaxis cases.

Methods: Using previously developed methods, we extracted features from unstructured narrative data using NLP and applied rule-based and similarity-based algorithms to identify anaphylaxis among 62 potential cases previously classified by human experts as anaphylaxis ($N = 33$), not anaphylaxis ($N = 27$), and unknown ($N = 2$).

Results: The rule-based and similarity-based approaches demonstrated almost equal performance (recall 100% vs 100%, precision 60.3% vs 57.4%, F-measure: 0.753 vs 0.729). Reasons for misclassification included the inability of the algorithms to make the same clinical judgments as human experts about the timing, severity, or presence of alternative explanations; and the identification of terms consistent with anaphylaxis but present in conditions other than anaphylaxis.

Conclusions: Although precision needs to be improved before these algorithms could be used without human review, we demonstrated that applying rule-based and similarity-based algorithms to unstructured narrative information from clinical records can be used for classification of anaphylaxis in the Sentinel System. Further development and assessment of these methods in the Sentinel System are warranted.

KEYWORDS

anaphylaxis, case classification, natural language processing, pharmacoepidemiology, sentinel system, validation

1 | INTRODUCTION

In May 2008, in response to the US Food and Drug Administration (FDA) Amendments Act of 2007, the FDA launched the Sentinel Initiative, a multi-year program for the establishment of a national electronic monitoring system for medical product safety.^{1,2} Funded under the Initiative, the Mini-Sentinel pilot was a collaborative effort between the FDA and a consortium of institutions to enhance the FDA's ability to actively monitor the safety of marketed medical products and complement the Agency's existing Adverse Event Reporting Systems. In 2016, the FDA announced the completion of the Mini-Sentinel pilot and the transition to the full Sentinel System.³ Through the Sentinel Common Data Model and automated analytical tools, the FDA can rapidly access and monitor information from electronic health records (EHRs) and insurance claims data in the Common Data Model, while safeguarding patient privacy.

Rapidly identifying and classifying health outcomes of interest (HOI) is a key component of Sentinel's active surveillance. Under the Mini-Sentinel pilot, FDA and Mini-Sentinel investigators initially evaluated computerized algorithms for classifying select HOIs by reviewing published literature to find algorithms that successfully used administrative data to classify these HOIs.⁴ Not all published algorithms were thought to be sufficient for the Mini-Sentinel purposes. Subsequently, several HOI algorithms were developed and validated using medical records in Mini-Sentinel, including one for anaphylaxis.^{5,6}

Improving the HOI algorithms for automatically classifying cases in EHRs is an important goal for the Sentinel System specifically and clinical informatics more generally. Most current approaches to classification of cases in EHRs are based on coded data, rule-based algorithms, and machine learning approaches.⁷ While traditionally used for information retrieval purposes, similarity-based methods can be used to identify documents (eg, narrative of an EHR) related to another document containing the key concepts of a case definition. Similarity-based methods rely on the identification of key terms in the document describing a clinical case and quantitatively comparing them with those same key terms in a case definition document. Similarity-based methods are very flexible because they only require the identification of the key medical terms, do not require coding of rules for case classification, and do not require a large training set for machine learning. We previously developed rule-based and similarity-based approaches to using unstructured narrative data for case classification.⁸⁻¹⁰ Using cases previously classified by human experts, we evaluated these approaches for application in the Sentinel System.

2 | METHODS

This section provides a description of the dataset, approaches to case classification and chart abstraction, and qualitative error analysis.

2.1 | Dataset

In the original study, Mini-Sentinel investigators selected a random sample of 150 patients who might be potential anaphylaxis cases

KEY POINTS

- Rule-based and similarity-based algorithms applied to unstructured narrative information extracted from medical records can be used for classification of anaphylaxis cases in the Sentinel System.
- Improvements in performance of the algorithms will depend on incorporating additional information that human experts use to distinguish anaphylaxis from other conditions.
- Further development and assessment of these methods in the Sentinel System is warranted to improve the efficiency and standardization of the case identification and classification process.

based on an electronic claims-based algorithm.⁶ Eligible patients must have at least 30 days of continuous enrollment in their respective health plan prior to the diagnosis date, which must occur between 1 January 2009 and 31 December 2010. An encounter was defined as a separate visit to a health care provider for which the chart materials include explicit documentation. For example, if a patient was seen in the emergency department (ED) and inpatient settings, and the chart included documentation from both settings, then that was counted as two encounters. However, since the goal of the original study was simply obtaining the information necessary to evaluate whether the case met the definition of anaphylaxis, if multiple eligible encounters occurred on the date of the outcome event, a single encounter could be selected. The duration of the encounter was the elapsed time in calendar days from the beginning of the first encounter through the end of the last visit (inclusive) documented in the chart. Medical charts from 122 of these patients were obtained by the investigators in the original Mini-Sentinel project, subsequently reviewed by a group of medical experts and classified for anaphylaxis based on the case definition proposed by Sampson et al.¹¹ Three labels were used for the chart classification: "yes" (ie, positive for anaphylaxis), "no" (ie, negative for anaphylaxis), and "unknown." The classification algorithms we used and described below were based on a case definition derived by the Brighton Collaboration from Sampson et al. We sought to reuse the labeled charts for this study.

The charts from the Mini-Sentinel anaphylaxis project were available in paper form, 60 handwritten and 62 typewritten. We did not have an efficient process for making the 60 handwritten charts machine readable, so we selected the 62 typewritten charts for the remainder of the study. For the 62 typewritten charts included in our study, the median number of encounters per chart was 1, and the mean was 1.39. The median time period of an encounter was 1 day, and the mean was 3.24 days.

We first used the ABBYY Fine Reader Optical Character Recognition software to digitize these 62 typewritten charts. The digitization process involved scanning, high-level error-checking, and editing of the digitized subset ($N = 62$) of the 122 medical records from the original study as well as the creation of a database of

electronic and searchable records. The 62 potential cases were previously classified by human experts as anaphylaxis ($N = 33$), not anaphylaxis ($N = 27$), and unknown ($N = 2$). The records included the redacted information for each patient chart organized by section; a set of dates (service start date, service end date, and extraction date); certain patient demographics; the adjudication forms used in the Mini-Sentinel project for the chart classification; and the final chart classification. For each chart, the ABBYY-scanned sections were categorized, consistently named, and attached in text format. For the purposes of the current study, we used the section text files, the final chart classification, and the adjudication forms from this database. The text files were organized into 12 different sections: allergies, clinic notes, discharge notes, ed forms, extraction forms, flowsheet data, follow-up plan, inpatient (IP) notes, labs, orders, visit notes, and vitals.

The reuse of the patient charts was approved by the data partners who participated in the original Mini-Sentinel anaphylaxis project. This original project was a collaboration among the FDA, the Mini-Sentinel Operations Center, and selected Mini-Sentinel Academic and Data Partners. Five Mini-Sentinel Data Partners participated in this project: (1) HealthCore, Inc.; (2) Humana; (3) three member health plans within the Kaiser Permanente Center for Effectiveness and Safety Research; (4) two member health plans within the Health Care Systems Research Network; and (5) Vanderbilt University (Tennessee Medicaid data). As illustrated in Figure 1, the text files were merged and then processed using two platforms: (1) the FDA's Event-based Text-mining of Health Electronic Records (ETHER) Natural Language Processing (NLP) system that extracts key information from clinical texts¹²; and (2) the National Library of Medicine (NLM) MetaMap tool that converts medical terms to the codes of selected terminologies, such as the

Medical Dictionary for Regulatory Activities (MedDRA).¹³ ETHER accomplishes feature extraction based on a set of rules. The primary and secondary diagnosis rules will label clinical terms as such when an appropriate trigger word is found within the parsed span of text, eg, "Dx" and "sDx," respectively. When a clinical term cannot be associated with any trigger word, it will be extracted as a symptom. Similarly, ETHER will retrieve clinical terms negated with the appropriate keywords, such as "without" and "denied." The full information for the ETHER algorithm can be found in our previously published work.^{9,10,12}

Following the removal of the negated terms, we used a local installation of the NLM MetaMap server and converted the ETHER output to MedDRA 18.1 terms. As MetaMap returns terms from any level of the MedDRA hierarchy, we had to further map them to the MedDRA Preferred Term (PT) Level. Each PT was linked to an ETHER diagnostic feature (primary diagnosis [PDx], secondary diagnosis [SDx], and symptoms [SYM]).¹² This supported the generation of a list of PTs (with their feature type) for each section per chart as well as a list of PTs (with their feature type) for each chart. The latter incorporated all the PTs from all sections.

2.2 | Approaches to case classification using unstructured narratives

We previously explored various approaches for the classification of spontaneous reports for anaphylaxis focusing on reports submitted to the FDA's Vaccine Adverse Event Reporting System.^{9,10,14,15} To classify the Mini-Sentinel charts in this study, we tested two approaches developed in our previous VAERS work. Each approach

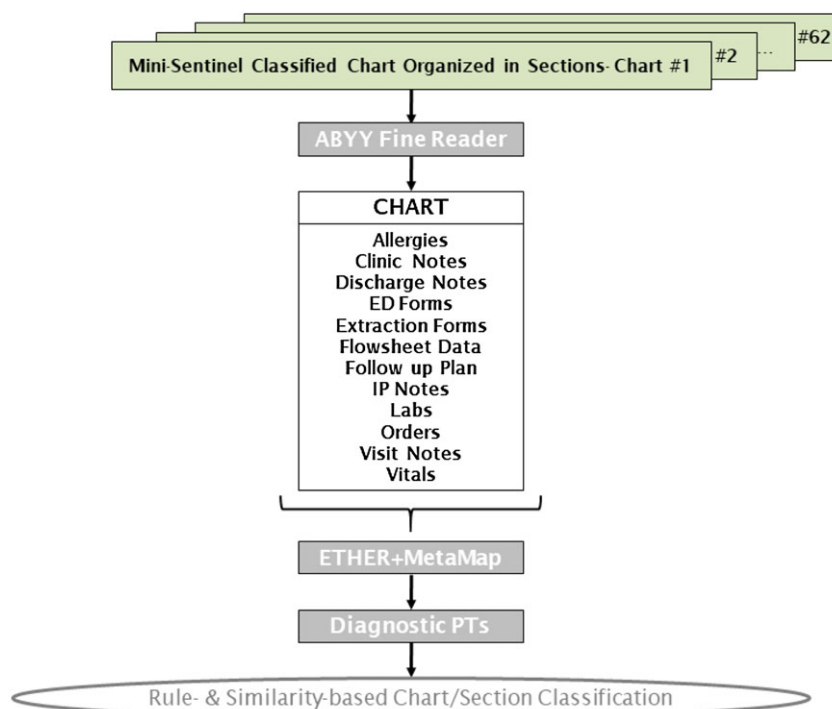


FIGURE 1 Text file processing—the text files were merged and then processed using two platforms: A, the FDA's event-based text-mining of health electronic records (ETHER) NLP system that extracts key information from clinical texts and B, the National Library of Medicine (NLM) MetaMap tool that converts medical terms to the codes of selected terminologies, such as the Medical Dictionary for Regulatory Activities (MedDRA) [Colour figure can be viewed at wileyonlinelibrary.com]

had two components: (1) an algorithm for using the medical terms to create a classification of either “yes” (ie, positive for anaphylaxis), or “no” (ie, negative for anaphylaxis); and (2) a set of medical terms that reflects the medical concepts in anaphylaxis. The approaches we used do not have an “unknown” category. We included the two “unknown” charts first with the negative charts and then with the positive for anaphylaxis to evaluate the impact on the overall performance of the classifiers.

The first approach, which we refer to as the “BC rule-based” approach, implements the modified Brighton Collaboration (BC) case definition for anaphylaxis,¹⁶ which is derived from and is highly similar to the Sampson et al case definition. The BC case definition for anaphylaxis derived from Sampson et al¹¹ uses similar medical concepts and terms and “... reflects a very similar understanding of anaphylaxis ...”¹⁶ as Sampson et al, but it allows for different levels of evidence and does not make assumptions about known allergens for the patient.¹⁶ One important difference between the Sampson et al definition used in the original Mini-Sentinel project, the original BC case definition, and the “BC rule-based” approach is that we included terms suggesting epinephrine use as a major criterion in our BC rule-based approach. This could result in a case classified as anaphylaxis with fewer clinical signs and symptoms present in the narrative compared with the number required in the Sampson et al definition or the original BC definition.

The second approach, which we refer to as the “enriched SMQ or eSMQ-based” approach, expands on the MedDRA PTs from the Standardized MedDRA Query (SMQ) for anaphylaxis with potential synonyms derived from the BC case definition.^{16,17} An information theoretic similarity calculation is applied to compare the PTs derived from the case with the PTs in the eSMQ to determine the classification.^{8,15} In this application, a similarity score > 0 leads to a classification of “yes,” and a similarity score = 0 leads to “no,” so that the presence of a single term from the eSMQ list in the terms extracted from a chart will result in a classification of “yes.” We anticipated that this strategy would result in high false positive rates but wanted to investigate the algorithms’ ability to identify the maximum number of true positive cases.

In addition to ETHER and MetaMap, the similarity-based approach was supported by the FDA’s Pattern-based and Advanced Network Analyzer for Clinical Evaluation and Assessment (PANACEA) that allows the identification of cases associated with selected conditions by performing the corresponding similarity calculations.¹²

For both approaches, the standard metrics of recall (ie, sensitivity), precision (ie, positive predictive value), and F-measure (harmonic mean of recall and precision) were calculated based on the actual classification of charts conducted in the Mini-Sentinel project⁶:

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

We also studied the distribution of the similarity scores calculated from the application of the eSMQ to the 62 charts. This included the

overall evaluation of the score distribution by calculating the similarity score between the charts and the eSMQ, counting the scores in a given bin, dividing the count by the total number of positive ($N = 33$) or negative charts ($N = 27$), and plotting and drawing a smoothed curve through the points. The scores for the two “unknown” charts were marked with an “X” or “+” depending on whether they were considered positive or negative, respectively. The similarity scores for the charts that did not contain any of the PTs found in the eSMQ were equal to zero and incorporated into the distribution. PANACEA was used for the calculation of the similarity scores, which further supported the construction of density plots using the corresponding function in R. We then calculated the classification performance for four similarity score thresholds: the minimum similarity score and the first, second and third quartiles of the score distribution.

2.3 | Contribution of chart sections to classification

We further investigated the particular contribution of the section-based and the various extracted feature types to the classification process. Each section inherited the label of the chart it belonged to, and the two approaches were applied to the section files and the lists of PTs as shown in Table 1.

2.4 | Qualitative error analysis (QEA)

For the rule-based approach, there are two potential sources of error. The first is if the algorithm incorrectly identifies a term in the chart as being the same as a term in the case definition (eg, the algorithm correctly finds a term [eg, abdominal pain], but the clinical context in the chart suggests it is not a persistent event as required by the case definition). The second is if the logic of the algorithm differs from the logic of the human expert in making the classification. In the similarity-based approach, the same first error can occur (ie, incorrect identification of a term in the chart as being the same as a term in the case definition). The second type of error can also occur but in a different manner: the similarity-based algorithm only requires the presence of terms in the chart and the case definition, without any particular logical ordering. Even a single term from the eSMQ list in the terms extracted from a chart will result in a classification of “yes.”

To assess these possible errors, we first established whether the terms in the case definition lists of the two approaches were found in misclassified charts. If they were found, then we evaluated whether the adjudicators had provided explanations about the logic they applied to the case.

3 | RESULTS

3.1 | Overall summary

The 62 potential cases were previously classified by human experts as anaphylaxis ($N = 33$), not anaphylaxis ($N = 27$), and unknown ($N = 2$). Both algorithms correctly classified all 33 of the human expert anaphylaxis cases as anaphylaxis, and the two unknown cases were also classified as anaphylaxis. Of the 27 cases classified as “not

TABLE 1 Application of the BC rule-based and the similarity-based approach to the chart and section files. The files belonging to the two charts labeled as “unknown” in the reference standard were considered as either positive or negative, and the classification results were calculated accordingly. In some cases, no false positives or false negatives were found, and precision and recall were not calculated (N/C)

			“Unknown” as Positive			“Unknown” as Negative		
			R	P	F-M	R	P	F-M
Rule based	Charts	62	100.0%	60.3%	0.753	100.0%	56.9%	0.725
	Allergies	17	27.3%	60.0%	0.375	30.0%	60.0%	0.400
	Clinic notes	51	92.9%	57.8%	0.712	92.3%	53.3%	0.676
	Discharge notes	22	100.0%	80.0%	0.889	100.0%	80.0%	0.889
	ED forms	21	100.0%	66.7%	0.800	100.0%	61.9%	0.765
	Extraction forms	52	0.0%	N/C	N/C	0.0%	N/C	N/C
	Flowsheet data	11	28.6%	66.7%	0.400	28.6%	66.7%	0.400
	Follow-up plan	1	N/C	N/C	N/C	N/C	N/C	N/C
	IP notes	1	N/C	N/C	N/C	N/C	N/C	N/C
	Labs	15	0.0%	N/C	N/C	0.0%	N/C	N/C
	Orders	31	42.9%	81.8%	0.563	45.0%	81.8%	0.581
	Visit notes	21	46.2%	60.0%	0.522	46.2%	60.0%	0.522
	Vitals	10	0.0%	N/C	N/C	0.0%	N/C	N/C
Similarity based	CHARTS	62	100.0%	57.4%	0.729	100.0%	54.1%	0.702
	Allergies	17	27.3%	42.9%	0.333	30.0%	42.9%	0.353
	Clinic notes	51	96.4%	55.1%	0.701	96.2%	51.0%	0.667
	Discharge notes	22	100.0%	72.7%	0.842	100.0%	72.7%	0.842
	ED forms	21	100.0%	66.7%	0.800	100.0%	61.9%	0.765
	Extraction forms	52	10.0%	N/C	N/C	3.6%	N/C	N/C
	Flowsheet data	11	28.6%	33.3%	0.308	28.6%	33.3%	0.308
	Follow up plan	1	N/C	N/C	N/C	N/C	N/C	N/C
	IP notes	1	N/C	N/C	N/C	N/C	N/C	N/C
	Labs	15	50.0%	N/C	N/C	44.4%	N/C	N/C
	Orders	31	76.2%	76.2%	0.762	80.0%	76.2%	0.780
	Visit notes	21	84.6%	61.1%	0.710	84.6%	61.1%	0.710
	Vitals	10	16.7%	100.0%	0.286	20.0%	100.0%	0.333

Abbreviations: BC, Brighton Collaboration; F-M, F-Measure; P, Precision; R, Recall.

anaphylaxis” by the human experts, 23 were misclassified as anaphylaxis for the BC rule-based approach, and 26 were misclassified as anaphylaxis for the eSMQ similarity-based approach. The results of the two classification approaches over the chart files and lists of PTs are shown in Table 1. The table illustrates that the rule-based and similarity-based approaches demonstrated almost equal performance (recall was equal to 100% in both approaches; precision for the BC rule-based vs the similarity-based classification with the two “unknown” charts classified as positive or negative: 60.3% vs 57.4% or 56.9% vs 54.1%, respectively).

In terms of the section classification (Table 1), although not all charts had all sections, it was shown that the classification algorithms all performed almost equally over the Clinic Notes section (recall for the BC rule-based vs the similarity-based classification with the two “unknown” charts classified as positive or negative: 92.9% vs 96.4% or 92.3% vs 96.2%, respectively; precision for the BC rule-based vs the similarity-based classification with the two “unknown” charts classified as positive or negative: 57.8% vs 55.1% or 53.3% vs 51.0%, respectively). In the case of the Orders section, the similarity-based classification was superior in terms of recall (for the BC rule-based vs the similarity-based with the two “unknown” charts classified as positive or negative: 42.9% vs 76.2% or 45.0% vs 80.0%, respectively), however inferior in terms of precision (precision for the BC rule-based vs the similarity-based with the two “unknown” charts classified as positive or negative: 81.8% vs 76.2% vs or 81.8% vs 76.2%, respectively). Some sections appeared in a small number of charts limiting the generalizability of conclusions about the contribution of these sections to classification. For

example, the “Discharge Notes” section seems to better support classification compared with other sections, but this section is present in only 22 of the 62 charts.

The similarity scores that were based on the similarity-based approach were used to plot the recall and precision values for the chart classification using the four similarity thresholds (Figure 2): the minimum non-zero similarity value and the first, second, and third quartile of the distribution. F-measure was maximized at the minimum similarity or the first quartile thresholds when the two “unknown” charts were classified as positive or negative. From the graph, one can observe that one “unknown” chart had a relatively low similarity score and the other relatively higher. The selection of the second and third quartiles reduced the overall performance in terms of recall and F-measure.

The distributions for the charts labeled as positive and negative in the reference standard are presented in the probability density plot that is embedded into the main plot shown in Figure 2. The relevance of the probability density plot is that it shows the distribution from which the first, second, and third quartile values are obtained, when calculating “Performance” in the larger Figure 2 graph. The interpretation of the probability density plot is that charts labeled as positive (green) received higher similarity scores versus the charts labeled as negative (red), so there is some discrimination provided by the terms in the eSMQ. However, one can observe that there is substantial overlap of the probability density functions suggesting that the terms in the eSMQ do not provide strong discrimination between positive and negative charts. In other words, the terms in the eSMQ do not uniquely define anaphylaxis cases.

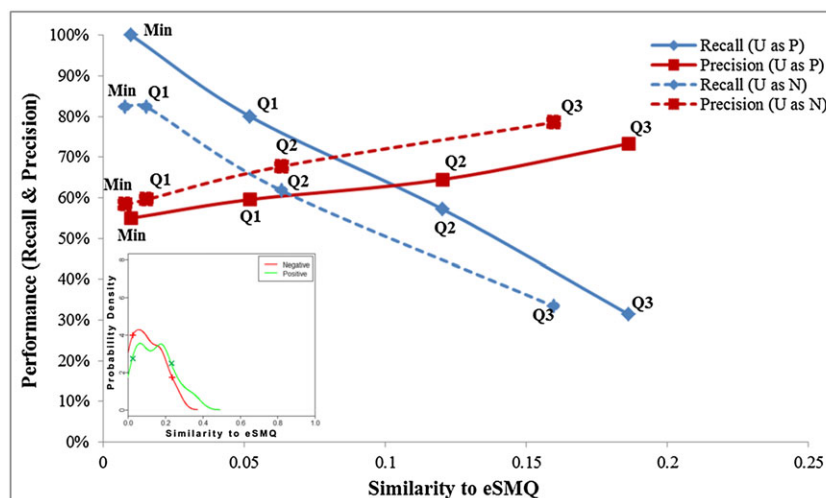


FIGURE 2 Recall and precision for the chart classification (all preferred terms used) based on the similarity-based and four different similarity thresholds. The thresholds were based on the minimum non-zero similarity value (min) and the first (Q1), second (Q2), and third (Q3) quartile of the similarity score distribution. Two pairs of lines are provided depending on whether the charts labeled as “unknown” in the reference standard were considered “positive” or “negative” in our analysis. The embedded probability density plot shows the similarity scores of the charts calculated in Pattern-based and Advanced Network Analyzer for Clinical Evaluation and Assessment (PANACEA). The scores for the chart labeled as “negative” and “positive” in the reference standards are represented by the red and green line, respectively; the two charts labeled as “unknown” were included in either the “positive” or the “negative” class, and the position of their scores on the corresponding line is marked with either the “X” or the “+” sign. eSMQ: Enriched Standardized MedDRA (Medical Dictionary for Regulatory Activities) Query; P: Positive; N: Negative; U: Unknown [Colour figure can be viewed at wileyonlinelibrary.com]

3.2 | Contribution of chart sections to classification

As described in the methods section, the text files were organized in 12 different sections. No chart included files for all of the 12 sections. The majority had more than half of the sections missing (mean number of sections $\pm$ standard deviation: 4 ± 2). Extraction forms, clinic notes, orders, discharge notes, visit notes, and ED forms were present in 52, 51, 31, 22, 21, and 21 charts, respectively; the remaining sections appeared in 15 or fewer charts (Table 1).

3.3 | Qualitative error analysis

All misclassifications were “false positives” (ie, charts misclassified as anaphylaxis when they were not classified as anaphylaxis by human experts). There were 23 misclassified cases for the BC rule-based approach and 26 misclassified cases for the eSMQ similarity-based approach, including the 23 misclassified cases for the BC rule-based, plus three additional cases.

We were able to find an explanation for the misclassification of 17/23 (74%) of the BC rule-based misclassified cases. For 13/23 (57%) of these cases, the BC rule-based algorithm identified the use of epinephrine, which was not part of the original classification criteria. However, the algorithm identified codes from at least two of the clinical areas (gastrointestinal, dermatological, respiratory) included in the criteria for all 13 cases. For one additional case, the original human adjudicator noted that the patient had received an Epi-pen so it was unclear if the case would have progressed to anaphylaxis. As a sensitivity analysis, we removed epinephrine as an indicator of anaphylaxis from the BC rule-based algorithm and did not observe an improvement in the classification (two anaphylaxis cases were incorrectly classified as “not anaphylaxis” and one previously misclassified

case was correctly classified as “not anaphylaxis.”) An additional 12 of the 23 (52%) misclassified cases (including nine of the cases for which the algorithm identified the use of epinephrine) had notations by the adjudicators that provided some indication of their clinical assessment. These notations included comments such as “hypotension 2 days later,” “O resp” (in a case the algorithm found a respiratory code), “? Persistent GI” (in a case where the algorithm found a gastrointestinal code), “facial angioedema only,” “not rapidly” and “4 days before wasp sting,” “had an abscess!,” “?corn syrup allergy?,” and “hives.” These notations suggest the human experts were making clinical judgments about the timing, severity, or presence of alternative explanations that the algorithm was not able to detect. For the remaining six misclassified cases, the algorithm correctly identified terms that fulfilled the logic of the algorithm, but the adjudication forms completed by the human experts did not indicate that the clinical signs and symptoms from the necessary systems (ie, dermatological, respiratory, cardiovascular, gastrointestinal) for anaphylaxis were present. We did not further review the original charts because this was beyond the scope of this study.

For the eSMQ similarity-based approach, we were able to find an explanation for 20/26 (77%) charts. The main reason for misclassification is that this approach only requires one anaphylaxis term to be considered an anaphylaxis case. One additional difference is that “epinephrine” was not a term used in the approach. In addition to the cases misclassified by the BC rule-based algorithm, the 3/26 (12%) additional charts that were misclassified by the eSMQ similarity-based algorithm contained terms consistent with anaphylaxis but also included many other terms suggesting the charts were describing patients with conditions other than anaphylaxis who had some of the same signs or symptoms.

4 | DISCUSSION

Although the false positive rate is relatively high, in this study we demonstrated that narrative information from medical records obtained from the Mini-Sentinel pilot can be used for classification of anaphylaxis reports when features are extracted using NLP techniques and automated coding strategies are applied to the extracted features. Low precision highlights the need for improving such algorithms before they are ready for routine use and incorporated into an EHR system. We also demonstrated that the use of an information theoretic similarity approach is as good as a rule-based approach for classification of anaphylaxis. These results should not be considered indicative of the general performance of the two methods as they might perform differently in the context of other phenotypes.

The sections of medical records contain information of different value for classification of anaphylaxis. Clinic notes, discharge notes, and ED forms had the highest recall, precision, and F-measures among the sections, although only a minority of charts had discharge notes and ED forms.

Strengths of this study include the case status determination was made by the human experts which serves as an excellent benchmark or "gold standard." The charts were obtained from a number of health systems, so our findings are potentially more generalizable than single-EHR studies. This study has some limitations. First, we re-used information from a previous validation study, and no additional data collection was conducted to support our analysis. For example, it would be insightful for our qualitative error analysis to know how the adjudicators used their clinical judgment in complex cases or resolved their disagreements. Second, we did not compare how our approaches worked compared with the human experts for all 122 cases because of the presence of many hand-written charts in the sample.

The algorithms themselves can be improved. We have applied classification approaches developed for post-market adverse event reports to EHRs without making any adjustments and achieved levels of precision similar to those in the original work. The misclassifications are all "false positives" which means one of the algorithms found terms consistent with anaphylaxis, when the human experts did not classify the chart as anaphylaxis. In the similarity-based approaches, a chart will be classified as anaphylaxis if it includes even a single term from the eSMQ case definition list. The similarity score will be low for a case with only one term, but it will be classified as positive. Identifying the optimal threshold similarity score might improve on this result. In the rule-based approach, we need involvement of at least two terms from two different organ systems or use of epinephrine to be classified as anaphylaxis. Removing use of epinephrine did not have a substantial impact on the classification. While the algorithms appeared to be appropriately identifying terms in the anaphylaxis definitions, the human expert's notations suggest the human experts were making clinical judgments about the timing, severity, or presence of alternative explanations that the algorithms were not able to detect. The development of more complex NLP strategies that incorporate context interpretation techniques may be able to better identify the alternative explanations found by the human experts. Such algorithms might be part of a second stage that are designed to address the specific issues that lead to false positive case classification. It is also

possible that the BC case definition itself is not sufficiently precise to lend itself to automation and improvement in the case definition is needed first. An iterative approach to improving the case definition as described by Walker et al¹⁸ might be worthwhile to explore. Any improvement in the case definition itself would likely need to be based on empirical evidence of what human experts include and do not include when classifying particular cases. More detailed annotations of case narratives, to include specific clinical judgments about cases, would help accomplish this. This would also likely inform improvement of the NLP algorithms.

Fully automated approaches to anaphylaxis case identification using clinical narratives may be possible. The current algorithms we tested have unacceptable precision, and improvements to the algorithms are needed before they could be used for automated extraction of cases. After algorithm refinement, prospective assessment of these methods on EHR data and paper medical records made machine readable through optical character recognition technology compared with claims algorithms currently in the distributed environment of the Sentinel System should be attempted.

ETHICS STATEMENT

The authors state that no ethical approval was needed.

ACKNOWLEDGEMENTS

This project and the original Mini-Sentinel anaphylaxis project were funded under the Mini-Sentinel task order HHSF22301012T from the US FDA. The authors would like to thank the data partners who contributed data to this project: Harvard Pilgrim Health Care, HealthCore, Inc., Humana, Inc., HealthPartners Institute, Kaiser Permanente Colorado, Kaiser Permanente Hawaii, Kaiser Permanente Northwest, and Vanderbilt University Medical Center/Tennessee Medicaid. We are indebted to the Tennessee Division of TennCare of the Department of Finance and Administration which provided data from the Tennessee Medicaid Program. The authors would also like to thank Aarthi Iyer and Susan Forrow for their help in the current project. Robert Ball and Taxiarchis Botsis are authors on US Patent 9,075,796, "Text mining for large medical text datasets and corresponding medical text classification using informative feature selection." For access to PANACEA and other algorithms used in this project please contact the FDA Technology Transfer Program at techtransfer@fda.hhs.gov. ETHER is available at <https://github.com/FDA/ETHER>.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Robert Ball  <http://orcid.org/0000-0002-1609-7420>

Sengwee Toh  <http://orcid.org/0000-0002-5160-0810>

REFERENCES

1. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the sentinel system—a national resource for evidence development. *N Engl J Med*. 2011;364(6):498-499.

2. FDA. FDA's Sentinel Initiative. Secondary FDA's Sentinel Initiative 2016. <http://www.fda.gov/Safety/FDASentinelInitiative/default.htm>
3. Ball R, Robb M, Anderson SA, Dal PG. The FDA's sentinel initiative—a comprehensive approach to medical product surveillance. *Clin Pharmacol Ther.* 2016;99(3):265-268.
4. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. *Pharmacoepidemiol Drug Saf.* 2012;21(Suppl 1):82-89.
5. Lanes S, Brown JS, Haynes K, Pollack MF, Walker AM. Identifying health outcomes in healthcare databases. *Pharmacoepidemiol Drug Saf.* 2015;24(10):1009-1016.
6. Walsh KE, Cutrona SL, Foy S, et al. Validation of anaphylaxis in the Food and Drug Administration's mini-sentinel. *Pharmacoepidemiol Drug Saf.* 2013;22(11):1205-1213.
7. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc.* 2016;23(5):1007-1015.
8. Cao H, Melton GB, Markatou M, Hripcsak G. Use abstracted patient-specific features to assist an information theoretic measurement to assess similarity between medical cases. *J Biomed Inf.* 2008;41(6):882-888.
9. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *J Am Med Inform Assoc.* 2011;18(5):631-638.
10. Botsis T, Nguyen MD, Woo EJ, Buttolph T, Winiecki S, Ball R. Vaccine Adverse Event Text Mining (VaeTM) system for extracting features from vaccine safety reports. *J Am Med Inform Assoc.* 2012;19(6):1011-1018.
11. Sampson HA, Munoz-Furlong A, Campbell RL, et al. Second symposium on the definition and management of anaphylaxis: summary report—second National Institute of Allergy and Infectious Disease/Food Allergy and Anaphylaxis Network symposium. *Ann Emerg Med.* 2006;47(4):373-380.
12. Botsis T, Jankosky C, Arya D, et al. Decision support environment for medical product safety surveillance. *J Biomed Inform.* 2016;64:354-362.
13. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc.* 2010;17(3):229-236.
14. Botsis T, Woo EJ, Ball R. The contribution of the vaccine adverse event text mining system to the classification of possible Guillain-Barre syndrome reports. *Appl Clin Inform.* 2013;4(1):88-99.
15. 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(7):573-582.
16. Rüggeberg JU, Gold MS, Bayas J-M, et al. Anaphylaxis: case definition and guidelines for data collection, analysis, and presentation of immunization safety data. *Vaccine.* 2007;25(31):5675-5684.
17. MedDRA Maintenance and Support Services Organization. Introductory guide to standardised MedDRA queries (SMQs) version 18.1. McLean, VA, September, 2015.
18. Walker AM, Zhou X, Ananthakrishnan AN, et al. Computer-assisted expert case definition in electronic health records. *Int J Med Inform.* 2016;86:62-70.

How to cite this article: Ball R, Toh S, Nolan J, Haynes K, Forshee R, Botsis T. Evaluating automated approaches to anaphylaxis case classification using unstructured data from the FDA Sentinel System. *Pharmacoepidemiol Drug Saf.* 2018;1-8. <https://doi.org/10.1002/pds.4645>