Can Natural Language Processing Improve the Efficiency of Vaccine Adverse Event Report Review?*

B. Baer1; M. Nguyen1; E. J. Woo1; S. Winiecki1; J. Scott1; D. Martin1; T. Botsis1,2; R. Ball1

1Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research (CBER), Food and Drug Administration (FDA), Rockville, Maryland, USA; 2Department of Computer Science, University of Tromsø, Tromsø, Norway

Keywords
Biosurveillance, product surveillance, post-marketing, data mining, natural language processing

Summary
Background: Individual case review of spontaneous adverse event (AE) reports remains a cornerstone of medical product safety surveillance for industry and regulators. Previously we developed the Vaccine Adverse Event Text Miner (VaeTM) to offer automated information extraction and potentially accelerate the evaluation of large volumes of unstructured data and facilitate signal detection.

Objective: To assess how the information extraction performed by VaeTM impacts the accuracy of a medical expert’s review of the vaccine adverse event report.

Methods: The “outcome of interest” (diagnosis, cause of death, second level diagnosis), “onset time,” and “alternative explanations” (drug, medical and family history) for the adverse event were extracted from 1000 reports from the Vaccine Adverse Event Reporting System (VAERS) using the VaeTM system. We compared the human interpretation, by medical experts, of the VaeTM extracted data with their interpretation of the traditional full text reports for these three variables. Two experienced clinicians alternatively reviewed text miner output and full text. A third clinician scored the match rate using a predefined algorithm; the proportion of matches and 95% confidence intervals (CI) were calculated. Review time per report was analyzed.

Results: Proportion of matches between the interpretation of the VaeTM extracted data, compared to the interpretation of the full text: 93% for outcome of interest (95% CI: 91–94%) and 78% for alternative explanation (95% CI: 75–81%). Extracted data on the time to onset was used in 14% of cases and was a match in 54% (95% CI: 46–63%) of those cases. When supported by structured time data from reports, the match for time to onset was 79% (95% CI: 76–81%). The extracted text averaged 136 (74%) fewer words, resulting in a mean reduction in review time of 50 (58%) seconds per report.

Conclusion: Despite a 74% reduction in words, the clinical conclusion from VaeTM extracted data agreed with the full text in 93% and 78% of reports for the outcome of interest and alternative explanation, respectively. The limited amount of extracted time interval data indicates the need for further development of this feature. VaeTM may improve review efficiency, but further study is needed to determine if this level of agreement is sufficient for routine use.

1. Introduction

In the United States, the post-market vaccine safety system relies in part on spontaneous reporting of adverse events by the public and health care providers. These reports are submitted to the Vaccine Adverse Event Reporting System (VAERS), which is managed jointly by the FDA and CDC. Other similar systems exist internationally, including the European EudraVigilance system and the WHO Uppsala Monitoring Center’s Vigibase. Although these systems have limitations for making causal inference about vaccines and adverse events [1], they continue to serve an important early warning function as was seen during the 2009 H1N1 influenza vaccine campaign [2]. These systems are used to detect signals of rare adverse events that might be associated with a vaccine. The best known example is the detection of intussusception after the Rotashield vaccine in 1998–1999 [3].

The review of spontaneous reports generally involves two complementary approaches. The traditional approach involves manual review of a series of case reports that have been submitted over a defined time period to identify unusual clinical and epidemiologic patterns [4, 5]. The case series remains largely dependent on human expertise with only limited automation. However, an increasing volume of reports associated with more and more medical product approvals may eventually overcome the human resources needed to review these reports manually. The second approach relies on statistical data mining of spontaneous reports based on medical product exposure-adverse event pairs, and is highly automated but requires extensive
human expert evaluation to interpret the analyses.

Both statistical data mining algorithms and individual case report review are routinely conducted for VAERS data [6, 7]. We previously developed the Vaccine adverse event Text Mining (VaeTM) system to systematically extract key features from the narrative of a report: clinical characteristics, time to onset since exposure to the vaccine, and alternative explanations for the adverse event [8]. This information is complemented by structured data such as the age, sex and vaccine name that is directly captured by the VAERS form.

To assess VaeTM, we evaluated the ability of the text mining and natural language processing system to effectively retrieve key clinical information from the narrative and present it to the end user (i.e. the safety evaluator) in a standard format. We set out to evaluate whether the VaeTM output could minimize the time and effort required to perform individual case review with minimal impact on the accuracy of the clinical interpretation.

2. Methods

2.1 VaeTM System

The VaeTM system processes the adverse event (AE) narrative reports and combines the text mining output with the structured information stored in VAERS. The VaeTM’s text mining module comprises multiple components that pre-process the free text, tag it using the VAERS dictionary and a set of regular expressions, parse it based on a set of grammar rules, and extract a full set of features (Table 1, “Full feature set” column). The feature extraction is mainly supported by the combined use of: 1) ‘triggers’, such as ‘hx’ and ‘dx’ for medical history and primary diagnosis, respectively; and 2) main and nested grammar rules. A certain form of sense disambiguation is performed by the incorporation of negation, possibility, impression, assessment and certainty information in the extracted features. A complete description and a thorough technical evaluation of the VaeTM system have been included in our previous publication [8].

The VaeTM system has been used in various studies. We have mapped the extracted medical terms to the criteria of the Brighton Collaboration case definitions of anaphylaxis [9] and Guillain-Barré Syndrome (GBS) [10] and classified the reports with varying levels of recall (83.1% and 60.5%, respectively); this difference was explained by the fact that the current VaeTM version does not extract laboratory information, which is necessary for the review of reports potentially associated with conditions like GBS [8, 11]. We further investigated whether the VaeTM output can feed other types of analyses and support other activities, such as automated knowledge generation from the medical literature [12]. These satellite efforts indicate the significance of using a system like VaeTM to support various surveillance tasks.

As previously described [8], the text miner can be programmed to extract from the narrative report either a full diagnostic feature set (Supplementary Figure 1a) or a limited diagnostic feature set (Supplementary Figure 1b) as outlined in Table 1. The extracted diagnosis, cause of death, and second level diagnosis (anything stated as “assessment”, “impression” or “possible diagnosis” – it does not have the value of the primary diagnosis but cannot be considered a symptom either) comprise the “outcome of interest” attribute. The drug, medical history, and family history contribute to the “alternative explanation” attribute. The remaining features (symptom, lot number, rule out diagnosis and vaccine) do not fall under a particular attribute but may add some value to the review process. We therefore considered them “irrelevant” and “potentially relevant” for the review process that was based on the “limited feature”- and the “full feature”-based interpretation, respectively.

Table 1. The extracted diagnosis, cause of death, and second level diagnosis (anything stated as “assessment”, “impression” or “possible diagnosis” – it does not have the value of the primary diagnosis but cannot be considered a symptom either) comprise the “outcome of interest” attribute. The drug, medical history, and family history contribute to the “alternative explanation” attribute. The remaining features (symptom, lot number, rule out diagnosis and vaccine) do not fall under a particular attribute but may add some value to the review process. We therefore considered them “irrelevant” and “potentially relevant” for the review process that was based on the “limited feature”- and the “full feature”-based interpretation, respectively.

2.2 Medical Evaluation

On September 18, 2012, we queried all 9,447 serious reports received between Jan 2008 and Dec 2011 from the United States from VAERS. Of these, 7,647 (81%) reports had the outcome of interest field populated (either with primary diagnosis, second level diagnosis, cause of death or a combination of these) after applying the text mining algorithm. From this group, 1000 reports were randomly selected for the analysis dataset. The reports were divided into two sets of 500 to be read and summarized by two experienced safety evaluators in a cross-over fashion. Safety evaluator A reviewed the full report of the first set of five hundred while safety evaluator B reviewed only the text mining extraction for the same set. For the second set of five hundred, the safety evaluators switched
We developed a data collection tool for the purposes of this study that contained the three different versions of the report (full report, limited feature set, and full feature set), a free text field for safety evaluators to summarize their interpretation, and a built-in timer to capture the review time. Each safety evaluator summarized his/her clinical impression including the primary diagnosis, alternative diagnoses, past medical history, and time to onset for each report and output version. The time between the vaccination and the start of the adverse event could be determined by using any information presented in the extracted text or in the structured fields from the VAERS reports. For reports in which multiple symptoms were described or symptoms evolved over time, the safety evaluators determined the onset time of the principal event or diagnosis.

When reviewing the text miner extraction (either the limited or the full feature set), the safety evaluator was blinded to the full report. The safety evaluator first read the limited feature set from the VaeTM output of the VAERS report. He/she briefly summarized the clinical adverse event and the time to onset of the adverse event (using the extracted text and/or structured fields of vaccination date and onset date). Next, the safety evaluator read the full feature set of the same VAERS report and viewed the summary for the limited feature extraction he/she had just created for that report. If there were no changes after reading the additional information presented in the full feature set, the previous summary was retained. Alternatively, if needed, the safety evaluator modified the summary with the new information from the full feature set.

A third safety evaluator compared the brief summaries of each of the three versions of text. The metrics of assessment were the conceptual match of the:
1. Outcome of interest – the primary diagnosis or adverse event in the report.
2. Onset interval (match within 12 hours for time to onset <1 day; same day for onset 1 day to <1 month; same week for onset 1 month to 1 year; same month for onset ≥1 year).
3. Alternative explanation – a potential cause of the adverse event other than the vaccine administration.

For example, a report may list a case of Guillain-Barré syndrome with symptoms of paresthesias and numbness in the feet beginning 10 days after a vaccination. The report might also note that the patient had...
a viral illness a week before the vaccine was received. In this case, Guillain-Barré syndrome would be the outcome of interest. The onset interval would be 10 days, and an alternative explanation for the adverse event of Guillain-Barré syndrome would be a preceding viral illness.

The three metrics were scored in a binary fashion, with 0 indicating an inadequate match and 1 indicating an acceptable, medically equivalent match. Medical equivalency was based on the adjudicating safety evaluator’s clinical judgment of a conceptual match of the diagnosis, seriousness, organ systems involved, localized versus systemic reaction, and possibility of alternative explanation for the adverse event. The clinical summary derived from the full text narrative was considered the reference standard. If the summary from the limited feature or full feature set included misleading alternative diagnoses not mentioned in the full text safety evaluator’s summary, then it would not be considered a match. A qualitative analysis by the third and also a fourth experienced safety evaluator not involved with the original case review was later performed to explore the reasons why certain cases did not match. The 74 cases in which the interpretation of the full feature did not match the interpretation of the full text for the outcome of interest were reviewed and the reason for mismatch was categorized as one of the following: 1) the VaeTM extraction did not include relevant information, 2) the VaeTM isolated information that had an altered meaning without supporting context, or 3) the safety evaluator omitted relevant clinical information in the event summary. A second set of 56 cases which contained additional clinical information in the extracted feature interpretation compared to the full text interpretation were further reviewed for future development of VaeTM.

We calculated improvement in review efficiency by calculating the number and percentage in word reduction between the free text, full features, and limited features and the review time for each approach.

### 3. Results

The outcome of interest, the alternative explanation, and the onset interval accounted for 29.1%, 7.2%, and 2.2% of the extracted information, respectively, in terms of word counts. The remaining information that was not interpreted by the medical experts was mainly related to the SYMPTOM feature and was equal to the 61.6% of the extracted text (symptom, lot number, rule out diagnosis and vaccine information that were equal to 58.4%, 0.2%, 0.4% and 2.5% of the total extracted information).

#### 3.1 Match of Outcome of Interest and Alternative Explanation

The proportion of matches between the interpretation of the VaeTM full feature extracted data compared to the interpretation of the full text was 93% (95% CI: 91–94%) for outcome of interest and 78% (95% CI: 75–81%) for alternative explanation. For the limited feature extracted text, the proportion of matches was 89% (95% CI: 87–91%) for the outcome of interest and 74% (95% CI: 71–77%) for the alternative explanation (Table 2). For the outcome of interest (attribute 1), there were significantly more full feature matches than limited feature matches (McNemar p < 0.0001). There was fair agreement between the limited feature and full feature matching (Cohen’s kappa = 0.67).

#### 3.2 Match of Onset Interval

Text-miner extracted data on the time-to-onset was used by the safety evaluators in 14% of cases; it matched the full text time-
to-onset in 54% (95% CI: 46–63%) of those cases. For the remaining 86% of cases, the information came from the structured time data. When supported by structured time data from reports, the match for time to onset was 79% for both the full feature and the limited feature extracted text. There was very high agreement between the limited feature and full feature matching with a kappa = 0.98.

3.3 Qualitative Analysis of Non-matching Cases

The qualitative analysis reviewed the reason why the 74 cases (7%) had a non-matching outcome of interest when the interpretation of the full feature extraction was compared to the interpretation of the full text. It was found that non-matching cases were primarily due to one of the following:

1) VaeTM not including relevant clinical information (13 of 74 cases [18%]) – Because VaeTM reduces the amount of information extracted from a report, it may omit relevant clinical information in a few cases. Thus, this result is expected to occur occasionally. For example, a full text report of accidental asphyxia indicated that the case was fatal by including a comment regarding the “autopsy report.” As a result, the VaeTM extraction did capture that this case involved a death.

2) VaeTM isolating information that had an altered meaning without supporting context (57 of 74 cases [77%]) – For example, a report of an infant with a fever and a possible viral infection after vaccine administration included the words “admitted for sepsis work up.” Even though in the full report it is clear that the sepsis evaluation was negative, the VaeTM full feature extraction included “sepsis” as a symptom. This word out of context led the safety evaluator to erroneously include sepsis as a medical condition in the full feature interpretation.

3) The safety evaluator omitting relevant clinical information in the event summary (4 of 74 cases [5%]) – As an example, in one case the safety evaluator for the full text listed “recurrent syncope” as the main adverse event when the full text report focused more on the patient’s abdominal pain and underlying enlarged kidneys. The full feature extraction included “fainting” but the second safety evaluator appropriately focused on the more significant events related to the renal symptoms and evaluation when writing the full feature extraction interpretation.

Additionally, another set of 56 cases with a non-matching score for the full feature interpretation compared to the full text interpretation for the alternative explanation were further reviewed because the extracted text interpretation included clinical information that was not present in the full text interpretation. A qualitative analysis of these cases using the same three categories as above showed that for the full feature extraction interpretation 38/56 cases (69%) were due to an altered meaning out of context (reason 2 above) while 17 cases (30%) were due to safety evaluator omissions (reason 3) and 1 case (2%) was due to VaeTM excluding relevant information (reason 1).

Table 3 Descriptive statistics based on the counts and review times per report

<table>
<thead>
<tr>
<th>Feature</th>
<th>FT</th>
<th>FF</th>
<th>LF</th>
<th>RT (Time)</th>
<th>FF RT (Time)</th>
<th>LF RT (Time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>183.8</td>
<td>47.6</td>
<td>17.6</td>
<td>0:01:26</td>
<td>0:00:36</td>
<td>0:00:27</td>
</tr>
<tr>
<td>Median</td>
<td>151.0</td>
<td>39.0</td>
<td>14.0</td>
<td>0:01:14</td>
<td>0:00:30</td>
<td>0:00:28</td>
</tr>
<tr>
<td>StDev</td>
<td>120.4</td>
<td>32.3</td>
<td>13.4</td>
<td>0:00:52</td>
<td>0:00:22</td>
<td>0:00:29</td>
</tr>
<tr>
<td>Max</td>
<td>663.0</td>
<td>206.0</td>
<td>107.0</td>
<td>0:07:55†</td>
<td>0:04:36†</td>
<td>0:04:51†</td>
</tr>
<tr>
<td>Min</td>
<td>14.0</td>
<td>2.0</td>
<td>1.0</td>
<td>0:00:03</td>
<td>0:00:04</td>
<td>0:00:02</td>
</tr>
</tbody>
</table>

FT: Free Text; FF: Full Features; LF: Limited Features; RT: Review Time; StDev: Standard Deviation
* word and other units separated by whitespace
** extracted features included words only separated by commas and ampersands
† hour:minute:second
‡ Single outlier of 5:53:16 excluded from calculations as due to safety evaluator not closing the case when completed.
§ Single outlier of 1:57:42 excluded from calculations as due to safety evaluator not closing the case when completed.

3.4 Improvement in Review Efficiency

Table 3 shows the information reduction translated into word (or token) counts along with the time spent to review a report based on the free text, the full features, and the limited features. The extracted full feature text averaged 136 (74%) fewer words than the full text narrative and resulted in a mean reduction in review time of 50 (58%) seconds per report.

4. Discussion

4.1 Scientific Approach

To our knowledge, this is the first evaluation of a text mining system that measures...
two important practical trade-offs of using a text mining system for medical product safety surveillance: review time and accuracy of clinical interpretation. VaetM was designed to target and extract the most relevant clinical information to support assessment and present it to the end user in a feature-based summarized version [8]. In our current evaluation, the clinical conclusion from VaetM full feature extracted data agreed with the full text narrative in 93% of reports for the outcome of interest with a 74% reduction in text and a 58% reduction in time of review. The alternative explanation from the VaetM full feature extraction agreed with the full text in 78% of reports. The VaetM full feature extracted text contained four features not included in the limited feature set (symptom, vaccine, vaccine lot number, and rule out diagnosis), yet there was only a small difference (83% compared to 79%) between the match rates of each extracted summary with the full text narrative summary. The qualitative analysis shed light on this surprising finding. While in many cases the additional features added important information to the overall clinical picture, in other cases, the additional information taken out of the context of the full report led the safety evaluator astray. Characterizing and quantifying these types of erroneous interpretations are important considerations in the overall evaluation of VaetM and its application.

The VaetM system attempts to remove the less relevant information from the manual review process and make the necessary clinical information available to the end users. Certain parts of the extracted information may still be less pertinent for certain surveillance tasks. It is a matter of medical judgment to identify and filter out this part and would be risky to automatically perform further reduction of potentially significant information. To address this point, we initially explored whether the “limited feature set” could include the most relevant information for the initial evaluation of reports – this would reduce the amount of information the safety evaluators had to review even more than 74%. We then used the “full feature set” to examine how “irrelevant” the remaining information (under symptom, lot number, rule out diagnosis and vaccine features) was. We found a marginal improvement of 4% in terms of the total matches for the outcome of interest and alternative explanation, which indicates that this subset cannot be ignored. As it was mentioned above, we previously mapped the VaetM output to existing case definitions for classification purposes and this could support the identification of a more focused subset of the extracted information. However, this would serve a particular use case and would limit the thorough evaluation of a dataset including complex and rare safety patterns.

With the design of this evaluation, the estimated reduction in review time for the full feature set might be smaller because the safety evaluator had a preview through the limited feature set prior to viewing the full feature set. Additionally, the review time should not be misconstrued to represent the amount of time FDA takes in its overall safety evaluation for a particular product as it does not take into account multiple other aspects of the analysis including data mining and integration with other sources of safety data.

Regarding the time to onset information, most of the information the safety evaluator chose to use was captured in the structured field. Extracted temporal information was most useful in cases where the onset was less than 24 hours after vaccination. Those situations often included words such as “immediately after” or a certain time period later that helped define the event within a one-day period. The system does not currently extract other temporal information from the narrative and does not perform any deep natural language processing to follow conditions that evolve over time. We are currently working on introducing a temporal information extractor to the VaetM system, as well as a natural language processing functionality that will allow the efficient resolution of these weaknesses.

With this study, we have presented a novel strategy to evaluate the benefits and costs from the application of a text mining system to clinical texts, which is a critical step towards its potential implementation. This is a major step that is largely underestimated in Clinical Decision Support and Natural Language Processing studies that are conducted in non-generalizable settings [13, 14]. Historically, systems have been evaluated based on their ability to identify the expected key entities in annotated clinical texts [15–21] or to support very specific tasks, such as the evaluation of the inappropriate emergency room use [22] and the automated identification of diagnosis and co-morbidity in clinical records [23]. Technical evaluation is important, but a medical evaluation is also essential to determine whether the loss of sentence structure and contextual narrative results in a different interpretation of the relationship between the medical product exposure and reported outcome. Indeed, we found that reviewers can evaluate the text miner output version of the adverse event reports with reasonable accuracy in less than half the time required under a traditional approach. We believe that our results support the practical utility of feature-based and information reduction approaches.

4.2 Limitations

This medical evaluation of the VaetM has several factors that must be considered when interpreting the results. The first is that we scored the VaetM extraction based on medical officer interpretations of the text miner’s output rather than directly on the output itself. This was a conscious decision to focus on the resulting interpretation rather than only the words in the output. If certain words or diagnoses are included in the output but their setting changes the interpreted meaning, the automated system is not as effective as it might appear if one only looks at the output. As described in the qualitative analysis section of the results, however, this approach increases vulnerability to human error and human variation.

Additionally, the evaluation does not distinguish between the match rate of high value reports such as striking case reports, positive rechallenge cases, completely novel suspected associations, or potential risks identified at the time of licensure. This evaluation also did not address whether the low 7% mismatch rate is sufficient for routine use or how it might affect the timely awareness of a new safety issue for a vac-
B. Baer et al.: Can Natural Language Processing Improve the Efficiency of Vaccine Adverse Event Report Review

5. Conclusion

We performed a medical evaluation of the three features of outcome of interest, time to onset, and alternative explanation obtained from the VaeTM extracted text and the full text narrative. We believe that our quantitative analysis yields promising results for the use of VaeTM in daily review to identify safety signals, but there are still some areas for further exploration. Even with a 74% reduction in words, the clinical conclusion from VaeTM full feature extracted text agreed with the full text in 93% and 78% of reports for the outcome of interest and alternative explanation, respectively. The time to onset attribute was found to be more dependent on the structured field data. VaeTM may be an important tool to improve review efficiency, but further study is needed to determine if this level of agreement is sufficient for routine use.

References