Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Generation of an annotated reference standard for vaccine adverse event reports

Matthew Foster [a,*], Abhishek Pandey [a], Kory Kreimeyer [a], Taxiarchis Botsis [a,b]

[a] FDA Center for Biologics Evaluation and Research, Office of Biostatistics and Epidemiology. 10903 New Hampshire Ave, Silver Spring, MD, United States
[b] The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, United States

ARTICLE INFO

Article history:
Received 28 February 2018
Received in revised form 8 May 2018
Accepted 21 May 2018
Available online xxxx

Keywords:
VAERS
NLP
Annotation
Corpus
Reference

ABSTRACT

As part of a collaborative project between the US Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention for the development of a web-based natural language processing (NLP) workbench, we created a corpus of 1000 Vaccine Adverse Event Reporting System (VAERS) reports annotated for 36,726 clinical features, 13,365 temporal features, and 22,395 clinical-temporal links. This paper describes the final corpus, as well as the methodology used to create it, so that clinical NLP researchers outside FDA can evaluate the utility of the corpus to aid their own work. The creation of this standard went through four phases: pre-training, pre-production, production-clinical feature annotation, and production-temporal annotation. The pre-production phase used a double annotation followed by adjudication strategy to refine and finalize the annotation model while the production phases followed a single annotation strategy to maximize the number of reports in the corpus. An analysis of 30 reports randomly selected as part of a quality control assessment yielded accuracies of 0.97, 0.96, and 0.83 for clinical features, temporal features, and clinical-temporal associations, respectively and speaks to the quality of the corpus.

1. Introduction

The US Food and Drug Administration (FDA) Vaccine Adverse Event Reporting System (VAERS) is a spontaneous reporting system in which pharmaceutical manufacturers, medical practitioners, and patients or their representatives submit data regarding the safety of vaccines. These data support important surveillance tasks including the examination of safety concerns related to marketed products, the evaluation of manufacturer compliance to reporting regulations, and multiple research activities both within and outside the FDA.

Within an individual VAERs report, the most important adverse events described in the free-text report narrative are coded using the Medical Dictionary for Regulatory Activities (MedDRA) and stored in structured fields [1]. While this coding process captures most adverse events, clinical information such as medical and family history, as well as most temporal information, is not captured from the narrative. This lack of information negatively impacts safety surveillance by requiring large investments in time and effort to manually review the clinical narrative and rule out alternative causes of adverse events following vaccination.

To better leverage information within the clinical narrative and reduce the manual time and effort required for safety surveillance tasks, we previously developed the Event-based Text-mining for Health Electronic Records (ETHER) tool [2]. ETHER utilizes a rule-based natural language processing (NLP) algorithm to extract both clinical and temporal features from VAERS reports, map extracted adverse events to MedDRA codes, and assign time stamps to each adverse event. As part of a recent project for the development of a web-based NLP workbench, we were tasked to create a publicly available set of annotated VAERS reports (Fig. 1) which will be used as a reference standard to improve our current NLP algorithm, as well as aid researchers in creating their own clinical NLP systems. In this paper, we outline the methodology used to create this reference standard and provide statistics on the finalized corpus.

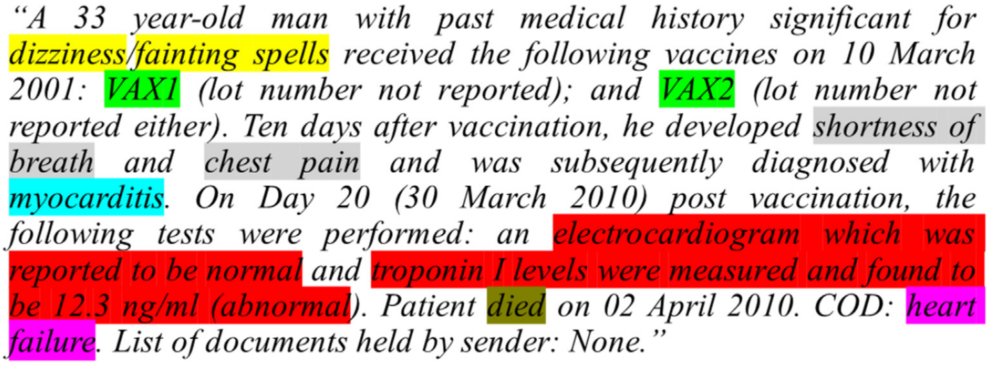
2. Methodology

The overall goal of our work was to create an annotated reference standard of VAERS reports for use in developing NLP systems. Before selecting a methodology to achieve this goal, we reviewed various annotation strategies reported in the literature as well as

* Corresponding author at: WO71 RM 1309-C, 10903 New Hampshire Ave., Silver Spring, MD 20993, United States.

E-mail address: matthew.foster@fda.hhs.gov (M. Foster).

https://doi.org/10.1016/j.vaccine.2018.05.079
0264-410X/© 2018 Elsevier Ltd. All rights reserved.



"A 33 year-old man with past medical history significant for dizziness/fainting spells received the following vaccines on 10 March 2001: VAX1 (lot number not reported); and VAX2 (lot number not reported either). Ten days after vaccination, he developed shortness of breath and chest pain and was subsequently diagnosed with myocarditis. On Day 20 (30 March 2010) post vaccination, the following tests were performed: an electrocardiogram which was reported to be normal and troponin I levels were measured and found to be 12.3 ng/ml (abnormal). Patient died on 02 April 2010. COD: heart failure. List of documents held by sender: None."

Fig. 1. Sample free-text narrative from a vaccine adverse event report annotated for clinical features. Features of interest such as the medical history (yellow), vaccine name (bright green), secondary diagnoses (gray), primary diagnoses (turquoise), laboratory findings (red), patient status (dark yellow), and cause of death (pink) have been highlighted. Annotated features are stored in a database which includes information such as the start and stop position of the annotated text, the feature type assigned to the text span, the report identification number the annotation is associated with, and a feature identification number which allows the database to keep track of the entry. The database contains annotations for multiple adverse event reports and this collection of annotations can be exported in various formats (e.g. CSV, XML) to assist users to in developing their own natural language processing systems. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

those used in major clinical NLP challenges [3–18]. A description of the findings of this review can be found in Appendix A. After conducting the review, we developed the annotation methodology outlined in Fig. 2, which breaks down the corpus creation process into four phases: (i) pre-training, (ii) pre-production, (iii) production-clinical feature annotation, and (iv) production-temporal annotation. This methodology was chosen to balance annotation quality and feasibility given the resource and time limitations mandated by our project.

The pre-training phase encompassed previous work by our group to create an annotation model for processing VAERS report features [19,20]. This annotation model defined nine types of clinical features (Primary Diagnosis, Secondary Diagnosis, Rule-out Diagnosis, Symptom, Cause of Death, Family History, Medical History, Drug Product, & Vaccine Product) and seven types of temporal feature (Date, Time, Duration, Relative, Age, Frequency, & Weekday). The rules for defining each feature type were developed with input from Medical Officers at the FDA and aim to capture specific aspects of the free-text narrative crucial for safety surveillance. The feature types in this annotation model served as the basis for the initial annotation model used in the pre-production phase.

In the pre-production phase, pairs of annotators applied the annotation model to identify either clinical features or temporal features and clinical-temporal associations. The annotations for each report were compared and inter annotator agreement (IAA) was calculated using average F-measure. F-measure was defined as: *2 * Precision * Recall/(Precision + Recall)*, which can be simplified to: *2 * True Positives/(2 * True Positives + False Positives + False Negatives)*. To calculate average F-measure we took the mean of two F-measure calculations performed by using each annotator's annotations as the gold-standard with which to evaluate the other's annotations.

For clinical features, temporal features, and clinical-temporal associations we calculated IAA using both strict and lenient matching. For clinical and temporal feature annotations, *strict matching* defined true positives as an exact match in feature text, start position, stop position, and tag type; while *lenient matching* considered true positives as annotations that captured the same concept but may have differed slightly in feature text, start position, stop position, or tag type. For clinical-temporal associations, strict matching defined true positives as clinical features associated with the same general temporal expression (i.e. considerable span overlap) AND the same association type category, as well as clinical features for which neither annotator assigned a temporal association. Lenient matches expanded on this definition but also included associations which fell under the *same expression* or *consistent timeline* mismatch categories.

In addition to calculating IAA, we also evaluated mismatched annotations as part of a qualitative mismatch analysis (QMA). The QMA involved a discussion amongst the annotators followed by the classification of mismatched annotations into various categories (Appendix B). If a consensus about which category the mismatched annotations belonged in could not be reached, an adjudicator was used to settle the disagreement. By allowing the quantification of various sources of disagreement, the QMA helped identify weaknesses in the annotation model which was then revised. Discussions amongst annotators as part of the QMA led to the incorporation of two new clinical feature types (Laboratory Findings & Patient Status), two new temporal feature types (Anchor & Other), and five relation type categories for clinical-temporal associations (Before, After, Overlap, Before-Overlap, and After-Overlap). These additions brought the final count of clinical and temporal features to eleven and nine, respectively. The finalized annotation guidelines, which define the various feature types and explain how each are assigned, can be found in Appendices C & D.

Throughout the pre-production phase, two types of report sets were used. *Training sets,* which consisted of less than 10 reports per set, were used to allow annotators to practice applying the annotation model and quickly resolve disagreements, while *testing sets,* which consisted of 30 reports per set, were used to evaluate whether annotators were applying the annotation model consistently over a larger number of reports. Once annotators exceeded specific IAA thresholds for clinical features (0.90), temporal features (0.80), and clinical-temporal associations (0.60), the annotation model was considered finalized and could not be changed.

The production phases involved the annotation of the 1000 randomly selected VAERS reports, which make up the final reference standard. The 1000 reports were split into two sets of 500 reports and each set was single annotated using a multi-pass annotation strategy. In the first pass of the production phase, the reports were annotated for clinical features (500 reports per annotator) while in the second pass of the production phase, these same reports were annotated for temporal features and clinical-temporal associations (500 reports per annotator).

Evaluation of the 1000 reports was conducted by review of 30 randomly selected reports (15 from each set of 500). For this evaluation, an adjudicator assessed the clinical features, temporal

Please cite this article in press as: Foster M et al. Generation of an annotated reference standard for vaccine adverse event reports. Vaccine (2018), https://doi.org/10.1016/j.vaccine.2018.05.079