

# Symptom Extraction and Linking from Vaccine Adverse Event Reports

**Thanapoom Phatthanaphan**

Stevens Institute of Technology  
tphattha@stevens.edu

## Abstract

The Vaccine Adverse Events Reporting System (VAERS) serves as a critical resource for monitoring the safety of vaccines. However, the vast amount of unstructured narrative text within VAERS reports poses a challenge for efficient analysis and symptom identification. In this paper, we present a novel approach to automatically extract and link vaccine-related symptoms from VAERS reports using sequence labeling techniques. Our method combines the power of named entity recognition (NER) and named entity linking (NEL) to achieve comprehensive and standardized symptom identification. We utilize NER packages and develop custom NEL methods to map identified terms to standard terminology from a symptom dictionary. By doing so, we aim to enhance the efficiency and accuracy of vaccine adverse event monitoring, ultimately contributing to the safety assessment of vaccines.

## Introduction

Vaccines are crucial for public health, and to ensure their safety, the Vaccine Adverse Events Reporting System (VAERS) was jointly established by the CDC and FDA. VAERS collects reports of adverse events following vaccination, allowing continuous safety monitoring.

However, VAERS faces a challenge due to the high volume of unstructured narrative reports it receives each year. These reports contain valuable information about adverse events, including symptoms. Identifying these symptoms promptly is essential for taking action and maintaining public trust in vaccination programs.

In recent years, sequence labeling techniques, used in linguistics and computational linguistics, have proven valuable for tasks like identifying named entities. Our project introduces an innovative method to automatically extract and connect vaccine-related symptoms from VAERS reports.

Our project focuses on two main goals: identifying symptom-related terms within the narrative text reports using NER tools and linking these terms to standardized symptom terminology. This approach addresses the challenge of unstructured VAERS reports while ensuring alignment with recognized medical terms.

This paper explains our methodology, including NER tool selection, symptom dictionary creation, and linking techniques. We also present experimental results to demonstrate the effectiveness of our approach. Ultimately, our project enhances vaccine adverse event monitoring, strengthens public health, and ensures vaccine safety, focusing on symptoms related to COVID-19.

Our project leverages innovative sequence labeling techniques, inspired by research in natural language processing. Specifically, the project draws insights from two prominent studies in the field. The work by Qi et al. [1] introduces a Python Natural Language Processing Toolkit, providing a valuable resource for processing diverse human languages. This toolkit is foundational to our approach, enabling efficient analysis of the narrative reports within VAERS. Additionally, the study by Zhang et al. [2] contributes essential biomedical and clinical English model packages for the Stanza Python NLP library. These packages are instrumental in our Named Entity Recognition (NER) methodology, enhancing the accuracy of symptom extraction from clinical narratives.

In alignment with the project's progression, we have successfully executed the critical phases of data preprocessing and the extraction of symptom entities associated with COVID-19 from the symptom text. Our next undertaking involves a meticulous comparison of the extracted symptoms with established standard symptoms. This comparative analysis aims to gauge the performance of our model, providing valuable insights into the accuracy and reliability of the implemented methodology.

While progress has been substantial, challenges may emerge during the subsequent phases of the project. Potential obstacles could include issues related to the diversity of symptom descriptions, variations in language use, and the inherent complexity of medical narratives. These challenges will be addressed through continuous refinement of the NER model, manual validation, and close collaboration with domain experts to ensure the accuracy and relevance of symptom extraction. Regular evaluations and iterative improvements will be integral to overcoming these challenges and ensuring the project's success.

## Problem formulation

The core machine learning task in this project is formally defined as a **multi-class classification** problem. This task involves classifying text descriptions of vaccine adverse events, specifically the SYMPTOM TEXT in the VAERS DATA table, into predefined categories representing various symptoms or symptom-related entities. The formulation can be represented as follows:

- **Input:** Text descriptions of vaccine adverse events (SYMPTOM TEXT).
- **Output:** Class labels representing symptom-related entities or symptoms.
- **Classes:** Multiple predefined classes representing different symptoms or symptom-related entities.

## Extraction of Symptom Entities using Stanza i2b2 NER Model

### Notations:

- $D$  represents the set of VAERS reports (Symptom text), each denoted as  $d_i$ .
- $S$  represents the set of symptom entities extracted from the reports using the Stanza NER package, represented as  $s_i$

The extraction process is denoted as  $s_i = nlp(d_i).entities$ , where  $s_i$  represents the extracted symptom entities.

## Methods

Several methodologies, techniques, and tools will be employed. These approaches aim to automatically identify symptoms from VAERS reports and link them to standard terms in a dictionary. Here's an outline of the key methods and tools required for this project:

### Data Preprocessing

Prior to symptom extraction, the reports in the VAERS dataset undergo a comprehensive data preprocessing phase. This step involves organizing the data for efficient processing. A dataset comprising 10,000 VAERS containing symptom text related to COVID-19 has been selected for comprehensive observation and testing, as illustrated in Figure 1. This sizable dataset is strategically chosen to evaluate and assess the performance of the employed model in the context of symptom extraction from VAERS reports. The selected dataset serves as a robust foundation for validating the efficacy and reliability of our methodology across a diverse and substantial set of real-world clinical narratives.

VAERS_ID		SYMPTOM_TEXT
0	2669769	body aches, fatigue Narrative: Took OTC Tyleno...
1	2527460	Headache, Myalgia, NauseaVomiting, chills Narr...
2	2673135	Headache, Fever, Body aches Narrative: Other ...
3	2672717	Headache & Myalgia Narrative: Other Relevant...
4	902418	Patient experienced mild numbness traveling fr...
...	...	...
9995	916173	REDNESS TO INJECTION SITE 12-30-20. PROGRESSED...
9996	916174	Patient described joint and muscle pain in the...
9997	916176	Numbness and tingling on left side of face, ey...
9998	916177	HIVES, tachypnea, vomiting - normal saline, ne...
9999	916178	Excessive swelling to left axillary lymph node...
10000 rows x 2 columns		

Figure 1: A dataset comprising 10,000 VAERS containing symptom text related to COVID-19

Before commencing the extraction of symptom-related entities from the set of 10,000 reports, The list of standard symptoms and the list of the most frequent 100 symptoms associated with COVID-19 are prepared. The purpose of these lists is to facilitate a rigorous comparison between the symptoms extracted from the reports and the established standard symptoms. Figure 2 provides an illustrative example of the list of standard symptoms.

```
['Fatigue',  
'Pain',  
'Chills',  
'Headache',  
'Myalgia',  
'Nausea',  
'Vomiting',  
'Pyrexia',  
'Hypoaesthesia',  
'Injection site hypoaesthesia',  
'Erythema',  
'Feeling hot',  
'Flushing',  
'Dizziness',  
'Electrocardiogram normal',  
'Hyperhidrosis',  
'Laboratory test normal',  
'Presyncope',
```

Figure 2: The list of standard symptoms

### Named Entity Recognition (NER) Packages

Utilization of specialized NER packages tailored for biomedical and clinical text analysis, including Stanza, i2b2, and similar libraries, to identify symptom-related entities.

This is the primary technique employed for symptom extraction. The i2b2 model is specifically chosen for its proficiency in recognizing clinical entities, aligning well with the medical context of VAERS reports.

### Symptom Extraction Implementation

The initiation of the Stanza library, employing the i2b2 Named Entity Recognition (NER) model, marks the commencement of the symptom extraction process from each individual report. The methodology includes an iterative procedure, wherein a subset of reports (specifically, 30 reports selected for the purpose of demonstration) is systematically processed. The output consists of the extracted entities, encompassing their respective text and assigned entity types. For clarity and illustrative purposes, Figure 3 serves as a representative example showcasing the extracted entities from the 30 reports.

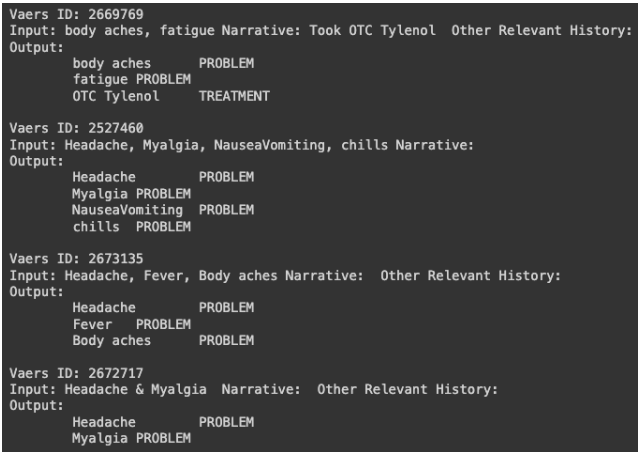


Figure 3: The example of extracted entities

### Named Entity Linking (NEL) Packages

The project involves the formulation and refinement of Named Entity Linking (NEL) methods, with the primary objective of establishing connections between symptom entities identified in the clinical text and standardized terms within a dedicated symptom dictionary. This phase of development is crucial for enhancing the interpretability and alignment of identified symptoms with recognized medical terminology, contributing to the overall robustness and applicability of the symptom extraction methodology.

### Rule-Based Matching

The project entails the creation and optimization of rule-based matching algorithms, encompassing both exact matching and fuzzy matching methodologies. These algorithms are designed to establish direct mappings of symptoms to standard terms by adhering to predefined rules. This

strategic approach aims to enhance the accuracy and efficiency of symptom mapping, providing a systematic framework for aligning identified symptoms with established standard terminology.

### Similarity-Based Matching

The project delves into an exploration of similarity-based matching techniques that leverage word embeddings, such as GloVe or clinical word embeddings. This investigation is undertaken with the aim of optimizing the accuracy of entity linking. The utilization of advanced word embeddings is envisioned to enhance the precision and efficacy of linking identified entities to standardized terms, contributing to the overall refinement of the research methodology.

### Challenges and Mitigation Strategies

- **Processing Time:** A potential challenge is the computational time required for extracting symptoms from a substantial number of reports. To address this, we employ a pragmatic approach by initially processing a limited subset (30 reports) for manual validation. This approach allows us to evaluate the correctness of the extraction method before scaling up to the entire dataset.
- **Manual Validation:** Given the complexity of medical narratives, manual validation is crucial for ensuring the accuracy of the extraction. While the manual checking of 30 reports may be time-consuming, it serves as a necessary step in refining the extraction methodology before applying it to a larger dataset.

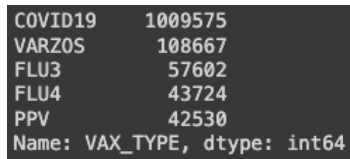
### Datasets and Experiments

The Vaccine Adverse Event Reporting System (VAERS) is a national early warning system to detect possible safety problems in U.S.-licensed vaccines. VAERS is co-managed by the Centers for Disease Control and Prevention (CDC) and the U.S. Food and Drug Administration (FDA). VAERS accepts and analyzes reports of adverse events (possible side effects) after a person has received a vaccination. There are three tables in the VAERS dataset that will be used in this project, including VAERS Data, VAERS Symptoms, and VAERS Vaccine.

- **VAERS Data:** The dataset contains detailed information about individual VAER submitted to the VAERS system.
- **VAERS Symptoms:** The dataset contains a list of symptoms and their corresponding codes that are reported in adverse event narratives.
- **VAERS Vaccine:** The dataset provides information about the vaccines including their characteristics and administration.

The datasets spanning from 2010 to the present day will serve as the foundation for creating new named entity recognition packages to identify terms related to symptoms and developing NEL methods to connect these identified terms to standardized terms present in a dictionary.

Due to a large dataset with more than 1 million rows, I will select a small sample based on one vaccine type by selecting from the vaccine type with the highest number of reports that is COVID-19 as shown in Figure 4.



COVID19	1009575
VARZOS	108667
FLU3	57602
FLU4	43724
PPV	42530
Name: VAX_TYPE, dtype: int64	

Figure 4: The top 5 vaccine types

The evaluation of the machine learning model's performance is a critical aspect of this project. However, it's essential to note that there is no ground truth annotation available for the data, making evaluation challenging. To overcome this limitation, a combination of automatic and manual evaluations will be employed.

- **Automatic Evaluation:** The model's accuracy in correctly classifying symptoms will be assessed through automatic evaluation metrics.
- **Manual Evaluation:** To ensure the accuracy of the model's predictions, a sample of clinical notes (typically 20~50) will be selected for manual evaluation. Experts manually review the results, verifying the correctness of the model's predictions.

The three key metrics, including precision, recall, and F1-score, will serve as essential tools for assessing the model's performance in identifying and linking symptoms.

- **Precision:** Precision will measure the accuracy of identified symptoms, ensuring their trustworthiness and relevance
- **Recall:** Recall will determine the model's ability to capture all relevant symptoms, minimizing the risk of missing significant health indicators
- **F1-score:** F1-score will provide an overall evaluation of the model's effectiveness in both accuracy and completeness.

## Project management

The project timeline and progress are outlined as follows:

### Completed processes:

- **Oct 10 – Oct 15:** Data Preprocessing.

- **Oct 16 – Oct 20:** Midterm Exams (Other courses).
- **Oct 21 – Oct 27:** Extracting Symptom-related Entities (Step 1).

### Ongoing processes

- **Oct 28 – Nov 2:** Link Entities to Standard Symptoms (Step 2).
- **Nov 3 – Nov 6:** Reporting Midterm progress
- **Nov 7 – Nov 12:** Continue to finish Step 2.
- **Nov 13 – Nov 17:** Evaluating the system.
- **Nov 18 – Nov 28:** Preparing presentation material.
- **Nov 29 – Dec 3:** Preparing Final report
- **Dec 4 – Dec 11:** Presentation period
- **Dec 11:** Submitting Final Report

## References

- [1] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, D.C. 2020. A Python Natural Language Processing Toolkit for Many Human Languages. The 58<sup>th</sup> Annual Meeting of the Association for Computational Linguistics. (pp. 101-108)
- [2] Zhang, Y., Zhang, Y., Qi, P., Manning, D.C., & Langlotz, P.C. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. Journal of the American Medical Informatics Association (Vol. 28, Issue 9). (pp. 1892-1899). <https://doi.org/10.1093/jamia/ocab090>