Symptom Extraction and Linking from Vaccine Adverse Event Reports

Name: Shweta Madhale

1. Introduction

Vaccinations have proven to be one of the most effective and impactful public health measures documented. These vaccinations can sometimes have possible post-vaccination adverse effects. This project aimed to automate the identification of adverse events and symptoms using a million Vaccine Adverse Event Reports (VAERS) with sequence labeling techniques. VAERs records post-vaccination events, and this project's key steps involved symptom entity extraction through Named Entity Recognition (NER) and linking these entities to standard symptom terms. Sequence labeling techniques are used to identify the symptom related entities in the VAERs reports.

This task was simplified due to some previously available work. With approximately a million records in the VAERS data, a subset was meticulously selected for processing. Yuhao Zhang and team's work on evaluating biomedical and clinical English text for NER outperformed scispaCy and BioBert in terms of performance and computational efficiency. Utilizing this model significantly enhanced the computational efficiency of the NER task. The works of Qingyu Chen et al and Zachary N. Flamholz et al, focused on generating word embeddings for clinical and biomedical text, had a crucial role in entity linking. The work of Jingcheng Du et al provided valuable insights into employing pretrained, domain-specific BERT for adverse events.

The data was loaded from VAERS report. The raw data was unsuitable for future tasks, so it was processed. A data subset was selected based on the vaccine type which was used for the initial step of NER. Later, more filtering was done by specifying the vaccine manufacturer. A thorough statistical analysis was performed to understand various trends pertaining to the data. The NER step made use of model from Yuhao Zhang et al. Symptom entities were procured after this, along with most common standard symptoms. The entity linking was performed using pre-trained clinical embeddings from Zachary N. Flamholz et al. The entity linking task was performed based on similarity matching with three models, namely Word2Vec, GloVe, and FastText on a word dimension of 100. The quantity of extracted symptoms was very large, although the models would be able to work on this number of symptoms, but it was tiresome for evaluation. So, the most common extracted symptoms were used for using on the models. All three models were used for the task of entity linking, and they were evaluated for the same. Evaluation was performed manually on small subset of data using a threshold on similarity score. Also, a ground truth was generated using ChatGPT for finding correct predictions. This project was able to investigate different existing models used for entity linking of clinical text.

There were some challenges encountered, like the diversity of unstructured language in VAERS reports. This challenge was overcome by robust preprocessing to standardize text. This included lowercasing, tokenizing, stemming, and removal of punctuation, special characters, and stop words. There was a need for meticulous data subset selection, which was done strategically by focusing on a vaccine type, and later the manufacturer. Addressing potential ambiguity in symptom mapping, similarity-based matching techniques use semantic similarities so can handle uncertain cases. While the absence of ground truth data posed a challenge for evaluation, ChatGPT was able

to map the symptoms intuitively. Despite these obstacles, the project laid a groundwork for more advancements in healthcare safety.

2. Problem Formulation

The process to extract and link symptoms was systematically framed within two key phases: an initial step focusing on symptom extraction and a subsequent stage dedicated to entity linking of the extracted symptoms to standard symptom terms. In the first phase, VAERS data served as the foundation, with emphasis on utilizing the 'Symptom Text' as the input to generate a comprehensive set of symptom entities as output. This extraction process essentially functioned as a specialized form of entity recognition, specifically targeting adverse events within the textual data. Moving on to the second stage, each extracted symptom entity underwent a critical process of linking to standard symptoms. Through symptom linking, these extracted symptoms were systematically mapped to standardized terms, utilizing a curated list of the most prevalent standard symptoms. Notably, the output of the initial extraction phase seamlessly fed into the subsequent linking phase, established a cohesive and iterative workflow.

The implementation strategy comprised of two primary tasks: Named Entity Recognition (NER) and Entity Linking. In the NER task, the primary objective was the identification and classification of named entities within the narrative text, with a focus on symptoms or problems as the designated named entities. This task resembled a binary classification task, wherein each word in the text was assigned a label indicating whether it represented a symptom or not. Transitioning to the Entity Linking task, the extracted symptoms were meticulously mapped to standard symptom terms, akin to solving a mapping problem where each symptom found its corresponding standard counterpart. The integration of these two tasks formed a holistic approach, initiating with a sequence labeling task where the 'Symptom Text' served as the sequence, followed by labeling with identified symptom and problem entities. This linking mechanism effectively transformed unstructured natural language data into a structured and meaningful list of symptoms.

The combination of symptom extraction and entity linking within this project not only facilitated the extraction of relevant information from unstructured textual data but also provided a systematic and organized representation of symptoms. The project was able to perform investigations about the pre-existing models for different stages of implementation effectively.

3. Methods

To achieve the project objectives of symptom extraction and linking from vaccine adverse event reports, different methods were used. The VAERS data had over a million records, so filtering of data was the initial step. The data was first filtered based on the vaccine type, ‘COVID19’ was selected, to get more insights by having more specificity, the data was further filtered based on manufacturer, ‘PFIZER\BIONTECH’. Also, the missing data was eliminated from selection. Based on the selection criteria, standard symptoms were obtained. A total of 14181 unique standard symptoms were obtained from the VAERS Symptom table. The symptoms were obtained by extracting symptom texts from the VAERS data table, cleaning the text (lowercase, punctuation removal, stop words, tokenization), and by selecting specific reports (around 10,000) based on VAERS_ID. This would be tiresome in later stages, so only the top 100 most common standard

symptoms were considered for further steps. Some formatting was performed on these symptoms to match the expected ground truth. The symptoms were lowercased, and underscores replaced blank spaces. A very thorough statistical analysis was conducted on this filtered data which helped understand different trends regarding demographics, adverse events, etc. In the symptom extraction stage, the ‘stanza’ package was used for NER. The pipeline included a ‘mimic’ package and ‘i2b2’ processor. Stanza has plenty of pre-trained models in the domain of biomedical and clinical texts, like the i2b2. i2b2 was used as it finds entities related to 'problem,' 'test,' and 'treatment'. This was decided based on the performance for work of Yuhao Zhang et al. Also, it allows fine-tuning the models to adapt to specific requirements with an exceptional performance. The symptom extraction returned symptom entities which were linked to standard symptoms in entity linking. The entity linking task was possible with rule-based matching, fuzzy matching, and similarity-based matching. For the problem in consideration, similarity-based matching approach was ideal as it captured semantic similarities between the symptoms. Also, similarity matching has a very flexible nature. The pre-trained clinical embeddings were used from the work of Zachary N. Flamholz et al. The models used for this were Word2Vec, GloVe, and FastText, with embedding of dimension 100. The count of unique extracted symptoms was more than 17000. The models were able to support this amount, but it was enormous for task of evaluation. So, only the 1500 topmost extracted symptoms were considered. The similarity matching was performed using cosine similarity. Using the similarity score as a threshold (0.6), the performance of the models was evaluated on a small subset of 50 linked entities. There was no ground truth available for this task, so ChatGPT was used for generating a ground truth. The most common standard symptoms were provided as a prompt along with small subsets of extracted top symptoms. This step caused the change for selecting only top extracted symptoms because of prompt and memory limits on ChatGPT. It generated the ground truth well, but there are limitations like lack of standard symptom, incorrect mapping, etc. So, the ground truth is not the absolute truth, but for this task it is acceptable. Using this ground truth, the correct predictions were counted for each model, and accuracy was obtained.

4. Datasets and Experiments

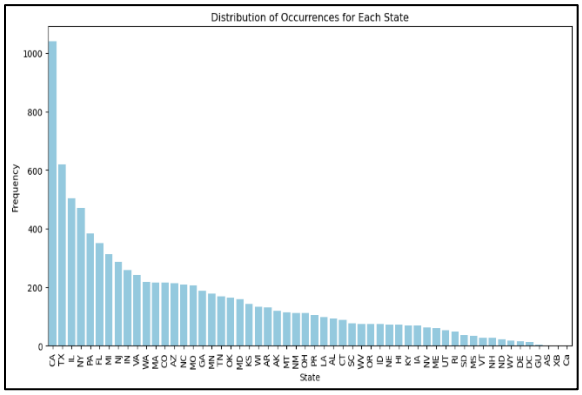
The dataset used was from Vaccine Adverse Events Reporting Systems (VAERS). This contained, the VAERS data table, VAERS symptom table, and VAERS vaccine table. The VAERS Data was used for symptom extraction with some preprocessing. The VAERS Symptoms was used to generate the standard symptoms list, and for symptom extraction and linking performance. The VAERS Vaccine filtered data, around 10000 reports, focusing on a specific vaccine type (COVID19) and vaccine manufacturer (PFIZER/BIONTECH). A thorough statistical analysis was performed for distributions. These observations were visualized as well.

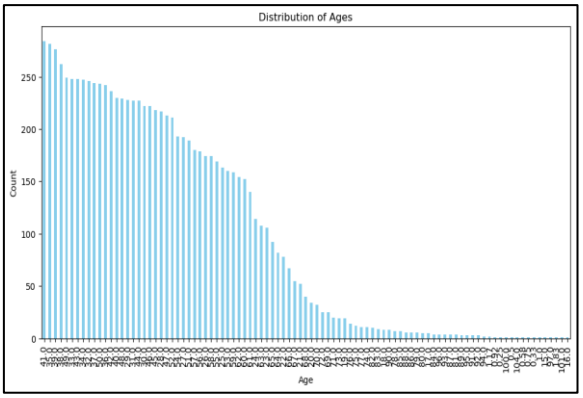
Parameter	**Analysis performed on COVID19 Data**
State Distribution	California (CA) had the highest turnout followed by Texas and New York. American Samoa, Guam, and Delaware had smallest turnout.
Age Distribution	The maximum count is for age 42(292) and minimum count for age 101(1)

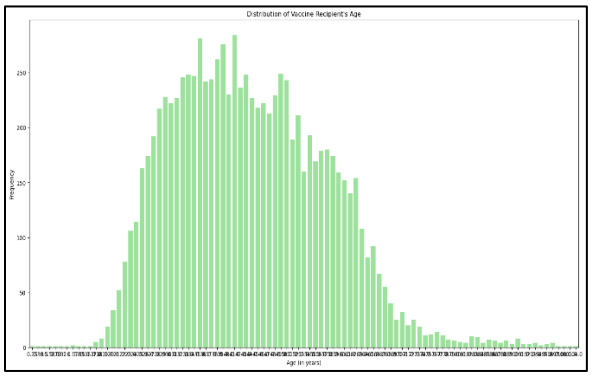
Gender	Vaccine administration for females was proportionally more than males and unknown
Symptom Text	The mean report length was around 375, minimum report length was 2, and maximum report length was 5904
Death	There were around 16 deaths reported as patient outcome against 9996 no death.
Life threat	There were around 61 life threatening patient outcomes against 9951 normal outcomes.
ER Visit	There were 18 reported ER visit patient outcome against 9994 normal ones.
Hospitalization	There were 135 reported hospitalization patient outcomes against 9877 normal ones.
Disability	There was a total of 10 reported disabilities as adverse event against 10002 normal ones.
Recoveries	There was a total of 4328 recoveries, 2842 no recoveries, and 2841 unknown.
Vaccine Manufacturer	Pfizer/BioNTech had a count of 7204
Vaccine Doses	The doses ranged from 1 to 7+, maximum count for 1 dose (8089) and minimum count for 6 doses (2)
Vaccine Route	The maximum count was Intramuscular (IM) with count of 7018 and minimum count of 8 for Intra Dermal (ID)
Vaccine Site	The maximum count was 5882 for LA and minimum count of 1 for RL
Vaccine Name	COVID19(PFIZER-BIONTECH) had 7204 reports

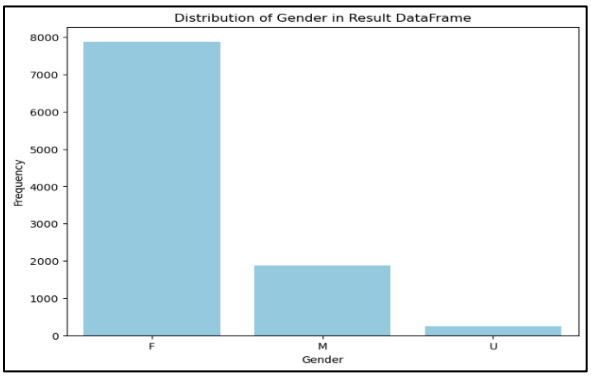
The statistical analysis of COVID-19 data reveals significant patterns and insights across various parameters. In terms of state distribution, California (CA) exhibited the highest turnout, closely followed by Texas and New York, while American Samoa, Guam, and Delaware had the smallest turnout. Analyzing age distribution, the data indicates a peak count for individuals aged 42 (292 reports) and a minimum count for those aged 101 (1 report). Gender-wise, vaccine administration was proportionally higher for females than males and cases where gender information was unknown. Examining symptom text, the mean report length stood at approximately 375 characters, with a minimum report length of 2 characters and a maximum report length of 5904 characters. Regarding patient outcomes, 16 deaths were reported against 9996 cases with no death, and 61 cases were deemed life-threatening against 9951 with normal outcomes. Emergency room visits, hospitalizations, and disabilities were reported in 18, 135, and 10 cases, respectively, against larger counts of normal outcomes. Recoveries totaled 4328, with 2842 cases showing no recovery and 2841 cases with an unknown recovery status. Pfizer/BioNTech emerged as the dominant vaccine manufacturer, with 7204 reports. Vaccine doses ranged from 1 to 7+, with the highest count for 1 dose (8089) and the lowest for 6 doses (2). Intramuscular (IM) was the most common vaccine

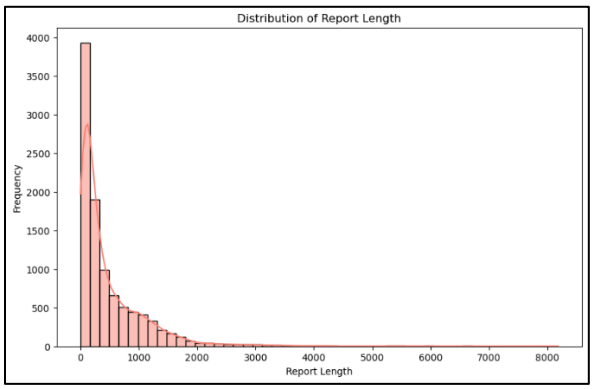
route with a count of 7018, while Intra Dermal (ID) had the minimum count at 8. Analyzing vaccine sites, LA recorded the highest count at 5882, while RL had the minimum count at 1. COVID-19 (Pfizer-BioNTech) had the highest representation among vaccine names with 7204 reports. This comprehensive analysis provides valuable insights into the distribution and outcomes associated with COVID-19 data, enabling a more informed understanding of the vaccination landscape.

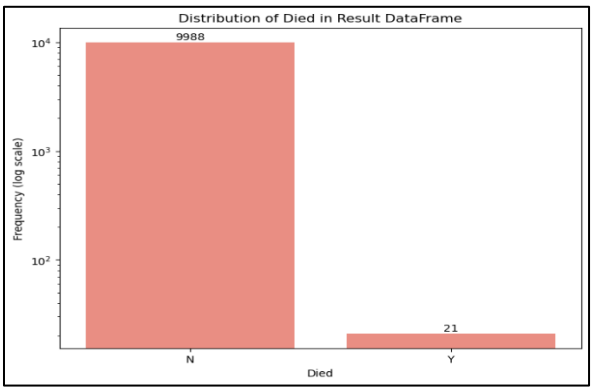


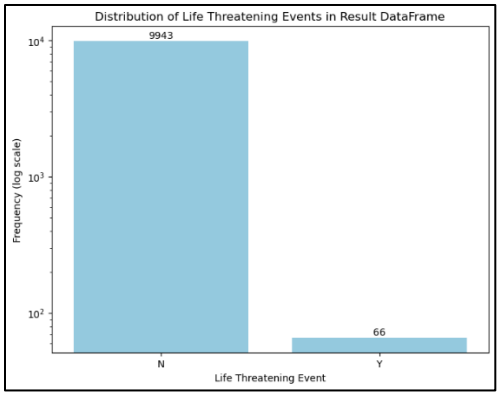


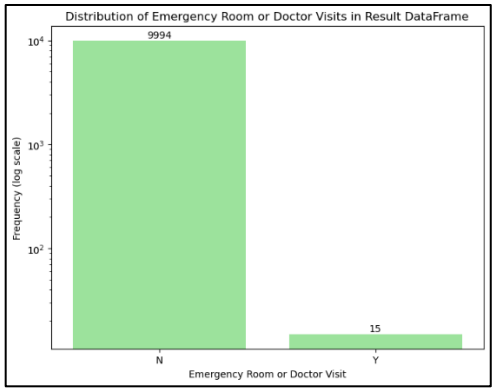


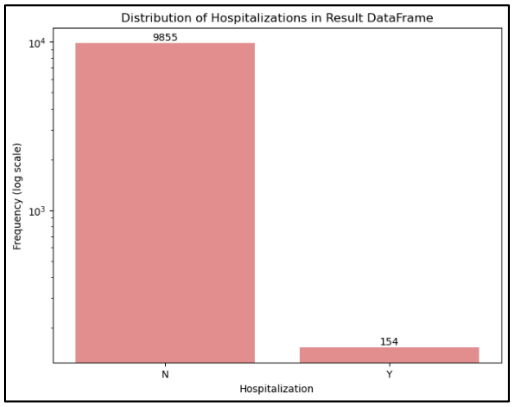


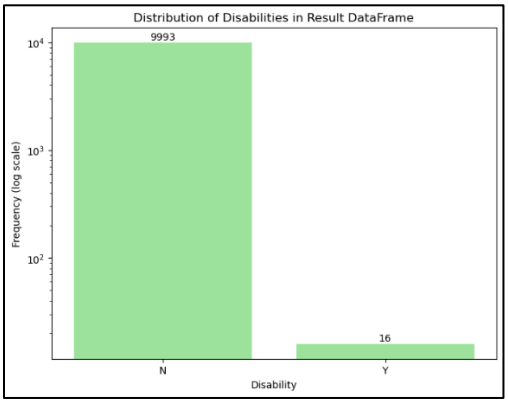


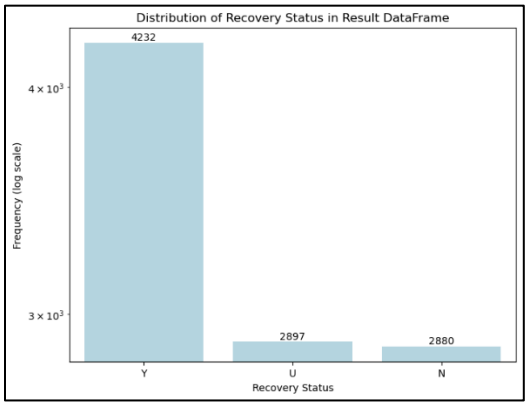


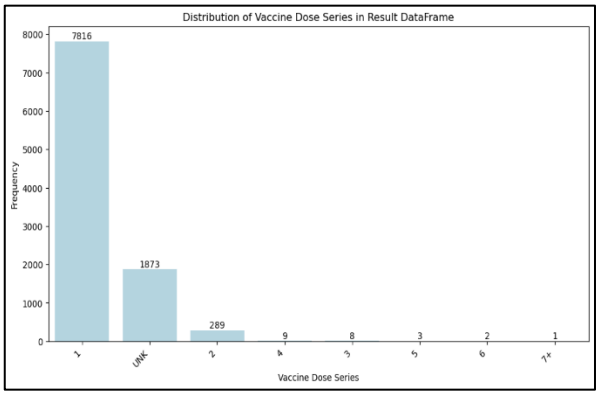

Distribution of Life Threatening Events in Result DataFrame
Frequency (log scale)
9943
66
N
Y
Life Threatening Event

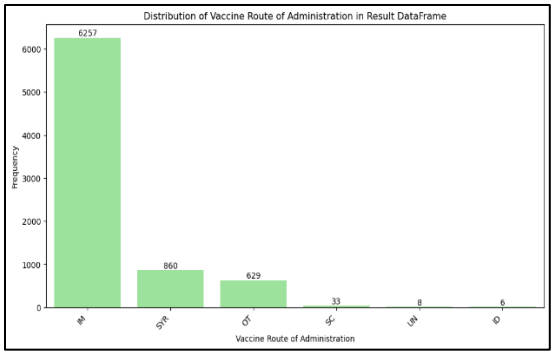

Distribution of Emergency Room or Doctor Visits in Result DataFrame
Frequency (log scale)
9994
15
N
Y
Emergency Room or Doctor Visit

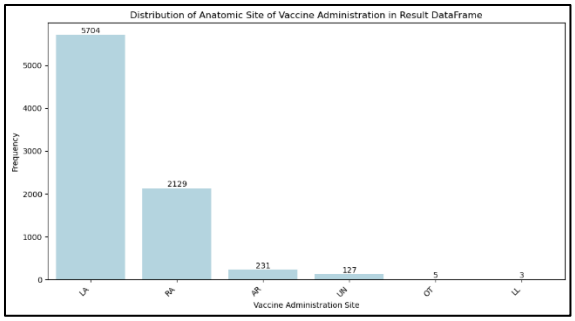

Distribution of Hospitalizations in Result DataFrame
Frequency (log scale)
9855
154
N
Y
Hospitalization


Distribution of Disabilities in Result DataFrame
Frequency (log scale)
9993
16
N
Y
Disability


Distribution of Recovery Status in Result DataFrame
Frequency (log scale)
4232
2897
2880
Y
U
N
Recovery Status


Distribution of Vaccine Dose Series in Result DataFrame
Frequency
7816
1873
289
9
8
3
2
1
1
UNK
2
4
3
5
6
7+
Vaccine Dose Series


Distribution of Vaccine Route of Administration in Result DataFrame
Frequency
6257
860
629
33
8
6
IM
SYR
OT
SC
UN
ID
Vaccine Route of Administration


Distribution of Anatomic Site of Vaccine Administration in Result DataFrame
Frequency
5704
2129
231
127
5
3
LA
RA
AR
UN
OT
LL
Vaccine Administration Site

The symptom text was lowercased, tokenized, stemmed, and cleaned. This was used as input for the NER process. The symptom extraction gave symptom entities. The symptom entities are to be linked to standard symptoms with similarity-based matching entity linking. Finally, metrics like precision, recall, F1 score, AUC, similarity scores will be obtained with corresponding objective functions.

With use of Symptom table, all symptoms were obtained, and then unique standard symptoms were generated. Using Symptom Text from the selected data, the data was processed, and this was used for symptom extraction. In symptom extraction, named entity recognition was used using stanza package, with mimic and i2b2 processor. The extracted symptoms were further processed by replacing blank space with underscore. From this, 17426 unique extracted symptoms were obtained. Out of this, the most common top 1500 extracted symptoms were selected from further process. Entity linking was performed using Word2Vec, GloVe, and FastText models. Iteration over extracted symptoms was done, and cosine similarity was computed with vector representation of extracted symptom and standard symptom. The standard symptom with highest similarity score was linked with the symptom. Using a threshold of 0.6 for similarity score, the models were evaluated on 50 linked entities. The word2vec model correctly classified 47, GloVe correctly classified 15, and FastText correctly classified 46. ChatGPT was used to generate ground truth by using standard symptoms and extracted symptoms. When the model performance was evaluated on this ground truth, the word2vec model had an accuracy of 19.11%, GloVe model had an accuracy of 14.12%, and FastText had an accuracy of 29.78%. This poor performance can be attributed to the meticulous selection of data in early steps, and an unreliable ground truth.

5. Conclusion

The project was investigative work on the existing models. There were two major steps involved in this. The first step included symptom extraction with named entity recognition, which performed similarly to a binary classification task. This step used the pre-existing stanza package with mimic and i2b2 processors. There is no ground truth for this step but based on manual evaluation on a small data subset, the stanza package performed well. In the second step, pre-trained clinical embeddings were used with Word2Vec, GloVe, and FastText models. The word embeddings were used to calculate cosine similarity and the highest score was used for linking. The ground truth was generated with ChatGPT, so is unreliable. Though the performance of the models was not good, it can be attributed to selectivity in the data, small amount of data used, unreliable ground truth, and insufficient standard symptoms. This project lays a good foundation for future works for similar projects. The parameters affecting the performance of the project can be worked upon to improve the performance, but work needs to be done to tackle the limitations of time and memory. In conclusion, while the investigative work on symptom extraction and linking using named entity recognition and pre-trained clinical embeddings lays a solid foundation, several challenges and opportunities for improvement are evident. The project, while providing valuable insights, serves as a starting point for continued exploration and refinement, necessitating ongoing evaluation, iteration, and adaptation to emerging medical knowledge and standards.

6. Project Management

Week	Task	Deadlines

1	• Defined project task. • Performed exhaustive research. • Decided implementation and methodologies.	
2	• Worked on project proposal. • Collected and loaded data	Project Proposal
3	• Performed Data Preprocessing • Selected model	
4	• Performed thorough statistical analysis	
5	• Implemented model for Extracting Symptoms • Work on midterm report	
6	• Optimize model for symptom extraction. • Evaluate performance of model	Midterm report
7	• Implement models for entity linking	
8	• Worked on presentation	
9	• Generate ground truth. • Evaluate performance	Presentation
10	• Final review	
11	• Worked on final report	Final Report and Code

7. Key references

1. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations 2020 Jul (pp. 101-108).
2. Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, Curtis P Langlotz, Biomedical and clinical English model packages for the Stanza Python NLP library, Journal of the American Medical Informatics Association, Volume 28, Issue 9, September 2021, Pages 1892–1899, https://doi.org/10.1093/jamia/ocab090
3. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. The 7th IEEE International Conference on Healthcare Informatics. 2019.
4. Zachary N. Flamholz, Lyle H. Ungar, Gary E. Weissmann Word embeddings trained on published case reports are lightweight, effective for clinical tasks, and free of protected health information, Dec 2019.
5. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021 Oct 15;3(1):1-23.
6. Jingcheng D, Yang X, Madhuri S, Meng Z, Jingqi W, Yuqi S, HuyAnh P, Hua X, Yong C, Cui T, Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning, Journal of the American Medical Informatics

Association, Volume 28, Issue 7, July 2021, Pages 1393–1400, https://doi.org/10.1093/jamia/ocab014
7. C Dreisbach, T Koleck, P Bourne, S Bakken, A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data, International Journal of Medical Informatics,Volume 125, 2019, Pages 37-46,
8. L Wu, F Petroni, M Josifoski, S Riedel, L Zettlemoyer, Scalable Zero-shot Entity Linking with Dense Entity Retrieval, https://doi.org/10.48550/arXiv.1911.03814
9. Z Nasar, S Jaffry, M Malik, Named Entity Recognition and Relation Extraction: State-of-the-Art, ACM Computing Surveys, Volume 54, January 2022.
10. VAERs Data - https://vaers.hhs.gov/data/datasets.html
11. Stanza Package - https://stanfordnlp.github.io/stanza/biomed.html
12. ChatGPT 3.5 - https://www.openai.com/
13. Clinical Embeddings - https://github.com/weissman-lab/clinical_embeddings/tree/master