ANALYZING TEMPORAL AND ASSOCIATIVE RELATIONSHIPS OF VACCINE ADVERSE EVENTS

CS-584
Prof. Ping Wang

Sneha Dharne
Gunik Luthra

Table Of Contents

Introduction

🔍 What is VAERS?

- **Definition:** Vaccine Adverse Event Reporting System (VAERS)
- **Purpose:** Tracks vaccine safety reports.
- **Scale:** Millions of records logged annually.
- Serves as a key tool for detecting potential safety signals.
- **Problem:** Unstructured text

Literature

- **Recent Advances in NLP for Biomedical Applications**
- **BERT & GPT-3:**
  - Effective in analyzing unstructured biomedical text.
  - Studies by **Chen et al. (2023)** demonstrate BERT's ability to extract adverse events from VAERS.
- **Integrating NLP with Statistical Models**
- **"Novel Data-Mining Methodologies" (2023):**
  - Combines Bayesian models with NLP for enhanced analysis of structured and unstructured data.
  - Foundation for hybrid approaches.
- **Focused Research on VAERS Data**
- **Zhang et al. (2019):**
  - Text mining for symptom identification.
  - Highlights the need for temporal analysis.
- **"Profiling COVID-19 Vaccine Adverse Events" (2023):**
  - Statistical + ontology-driven methods for symptom categorization.
  - Does not address symptom progression.
- **"COVID Vaccine and Cardiovascular Risks" (2023):**
  - Uses NLP for extracting cardiovascular risks.
  - Lacks temporal and associative insights.
- **Research Gap:**
- Need for hybrid models integrating **LLMs + statistical analysis** to capture temporal progression and co-occurrence of symptoms.

Task Overview

- For this project we aim to compare the analysis of temporal relationship and association relationship of symptoms extracted by a NER and a LLM.
- The analysis will be compared on various metrics such as:
  - Temporal Analysis:
    - Kendal Tau Rank Correlation Coefficient
    - Longest Common Subsequence
    - Dynamic Time Warping
  - Association Analysis
    - Method: Apriori Algorithm
    - Evaluation:
      - Support
      - Confidence
      - lift

Challenges

- **Unstructured Free-Text Data** 📝
- Difficult to standardize and analyze.
- **Medical Synonyms & Abbreviations** 🔤
- Variability in terminology creates inconsistencies
- **Scalability Issues**💻
- High computation time and cost for big data.
- **Limitations of NER Models** 🧠
- Struggles with semantic understanding (e.g., differentiating symptoms from medical history).
- **Annotation Complexity** ⏳✏️
- Manual labeling is time-consuming and challenging for non-medical professionals.

Solutions

Data Pre-processing

- VAERS DATASET:
  - 2024VAERSDATA.csv
    - VAERS_ID
    - SYMPTOM_TEXT
  - 2024VAERSVAX.csv
    - VAERS_ID
    - VAX_TYPE
  - 2024VAERSSYMPTOMS.csv
    - VAERS_ID
    - SYMPTOM1
    - SYMPTOM2
    - SYMPTOM3
    - SYMPTOM4
    - SYMPTOM5

DATA FOR TEMPORAL ANALYSIS

DATA FOR ASSOCIATIVE ANALYSIS

Experiments

Setting up the models.

🧹 Tokenization based on basic data cleaning and removal of stop-words

- Challenge: Medical abbreviation and a lot of text due to free text being a narration

🤖 NER based model, along with medical abbreviations handled: great at capturing symptom text

- Challenge: Captured tokens from medical history as symptoms, lacked semantic understanding of the free text, also captured synonyms as it is (body ache and body pain are considered two different kind of symptoms)

🚀 GPT-4o: great at enlisting symptoms.

- Challenge: Token limit, Difficult to scale for big data

🚀 Gemini 1.5 Flash

Hybrid Approach – A comparison

NLP Techniques

- Tokenize
- Lowercase
- Stop-word removal
- Medical Abbreviation dictionary
- en_ner_bc5cdr_md
- De-duping symptoms from symptom list

LLM Implementation

- Prompt Engineering
- One-shot Learning
- Chunking
- Optimizing token limit
- Formatting on the fly

Symptom extraction through NER and EDA

- Initially, symptoms were extracted by applying Named Entity Recognition (NER) techniques(en_ner_bc5cdr_md model), which involved removing stop words and ignoring abbreviations. To gain insight into the extracted symptoms, exploratory data analysis (EDA) was performed.

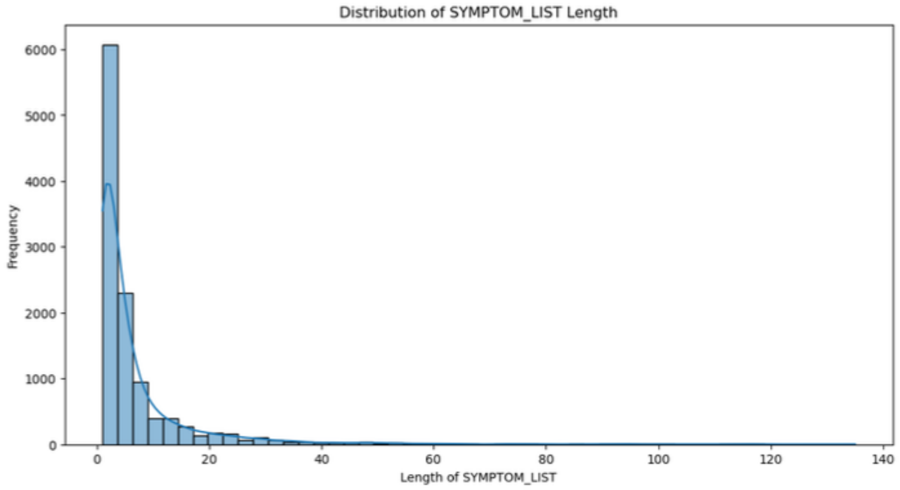
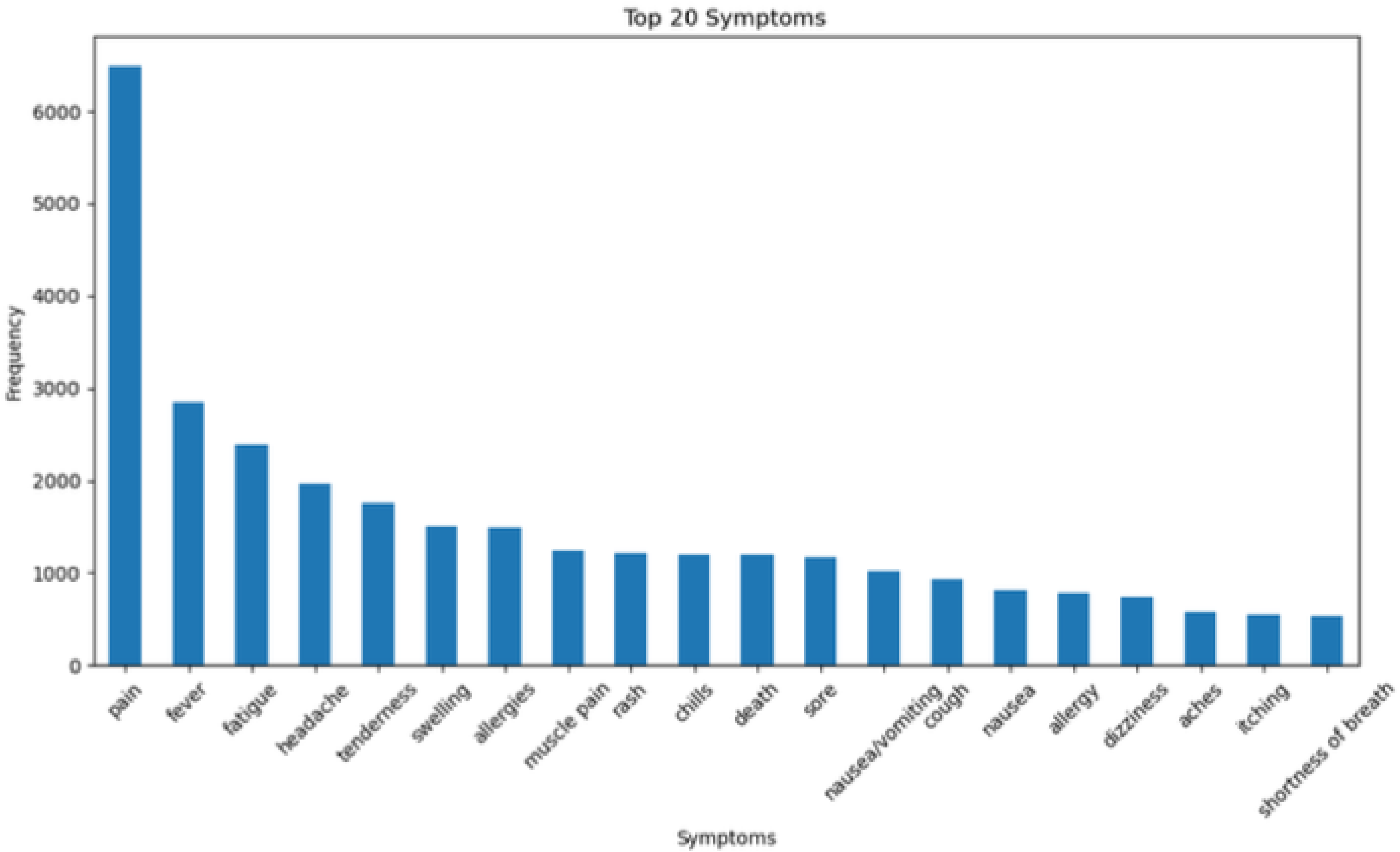
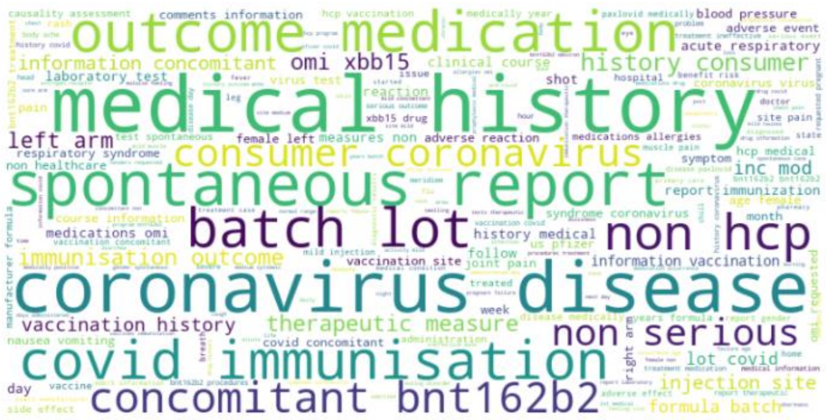
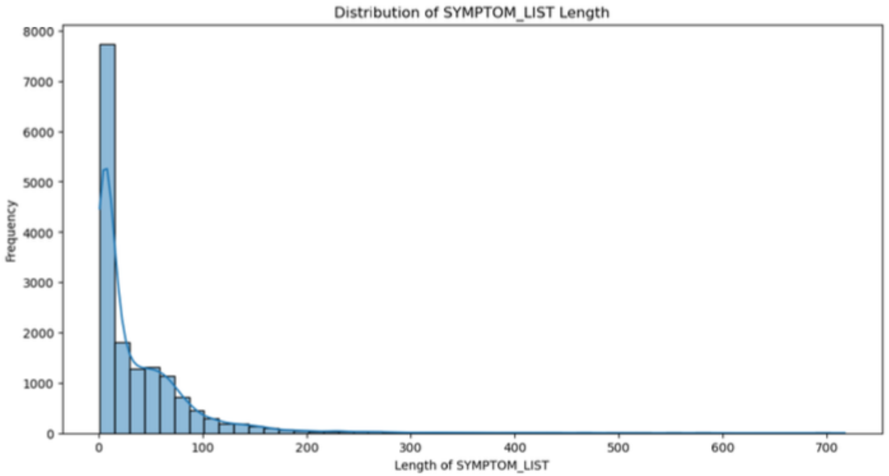


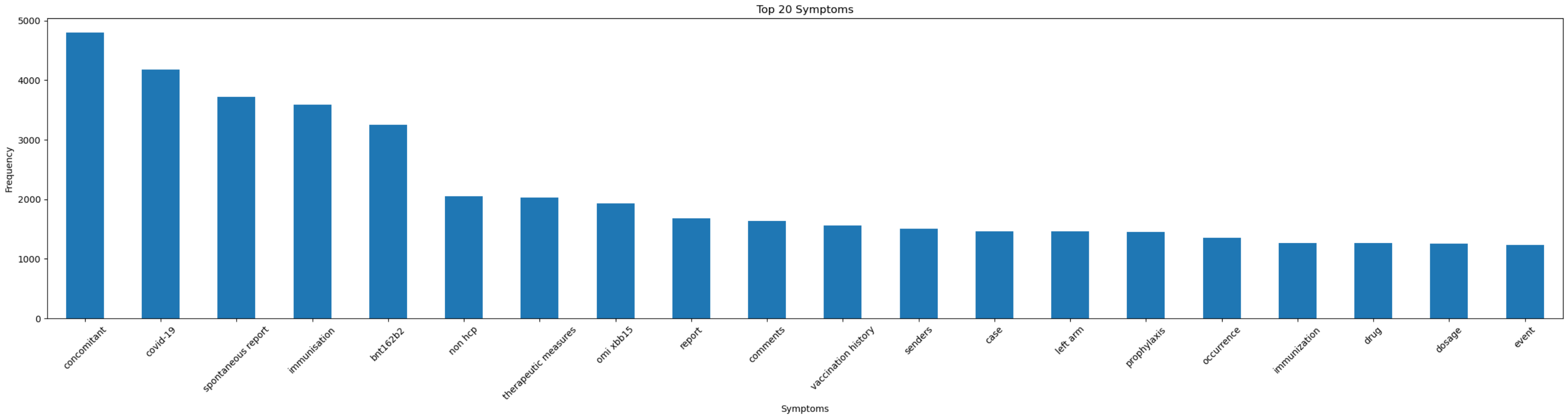
Figure 2: Number of Symptoms per data point


Top 20 Symptoms
Frequency
0
1000
2000
3000
4000
5000
6000
Symptoms
pain
fever
fatigue
headache
tenderness
swelling
allergies
muscle pain
rash
chills
death
sore
nausea/vomiting
cough
nausea
allergy
dizziness
aches
itching
shortness of breath

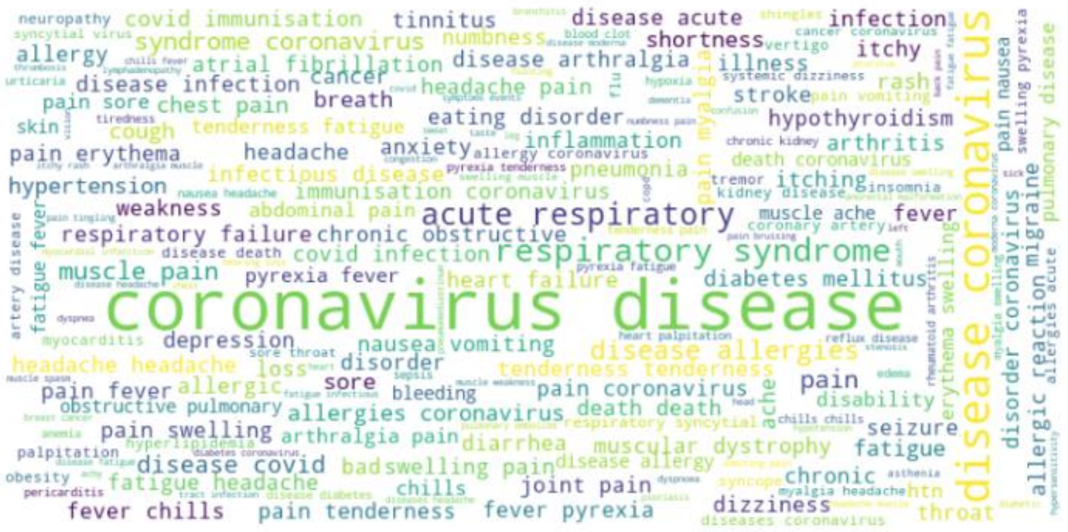
- Symptoms were extracted using NER techniques(en_core_sci_md) with removing STOP WORDS and replacing abbreviations(HR ,N/V, BP, etc) with their meanings. To gain insights EDA was performed.

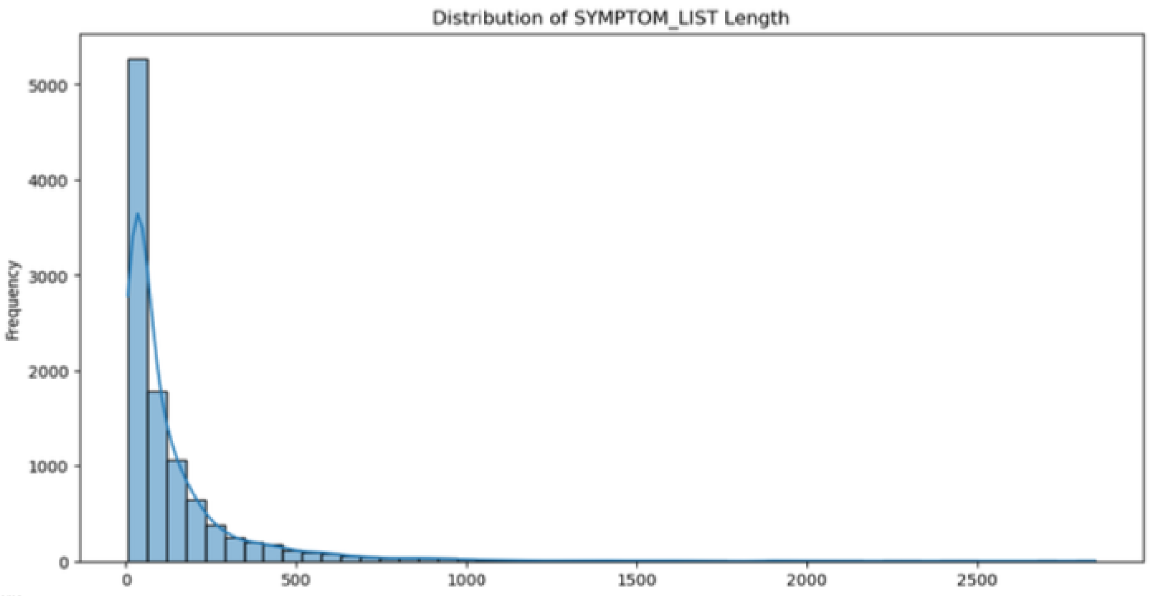


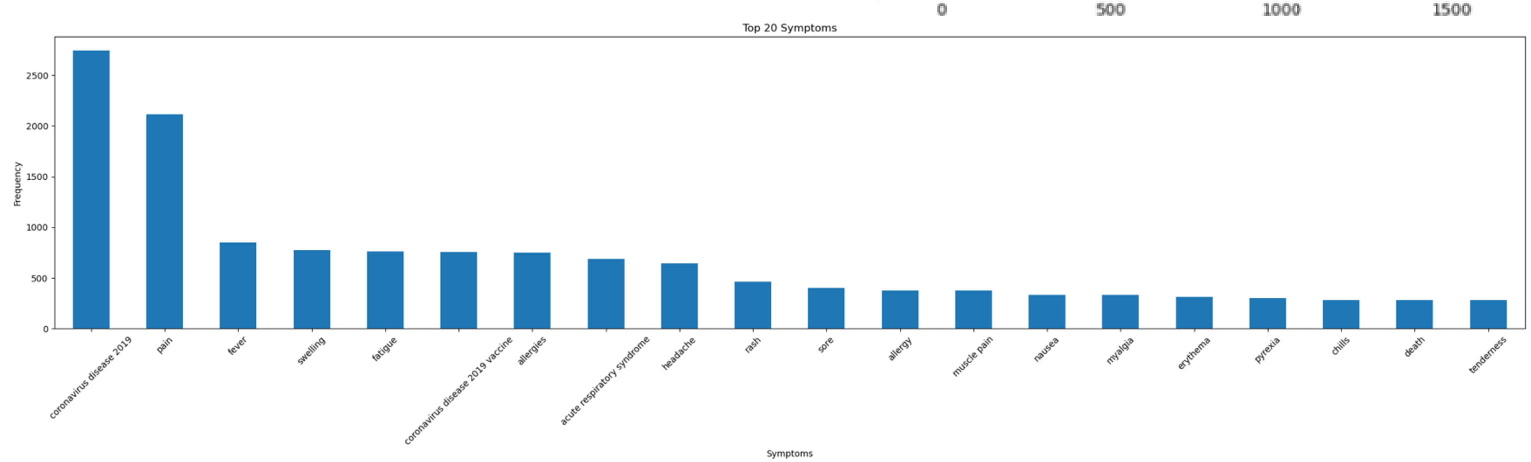




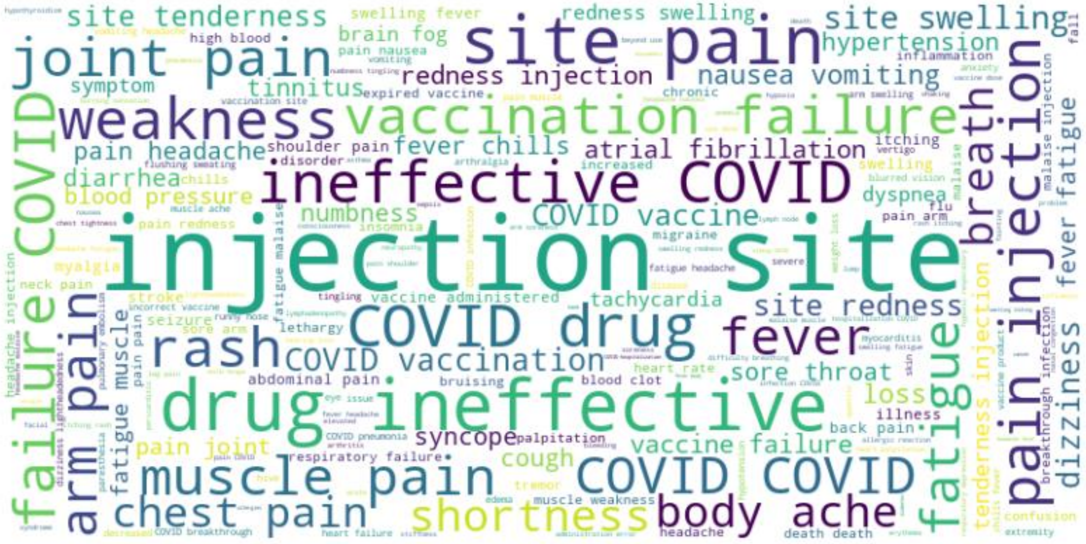
- Symptoms were extracted using NER Techniques(en_ner_bc5cdr_md model) removing STOP WORDS and replacing abbreviations(HR ,N/V , BP, etc) with their meanings. To gain insights EDA was performed.

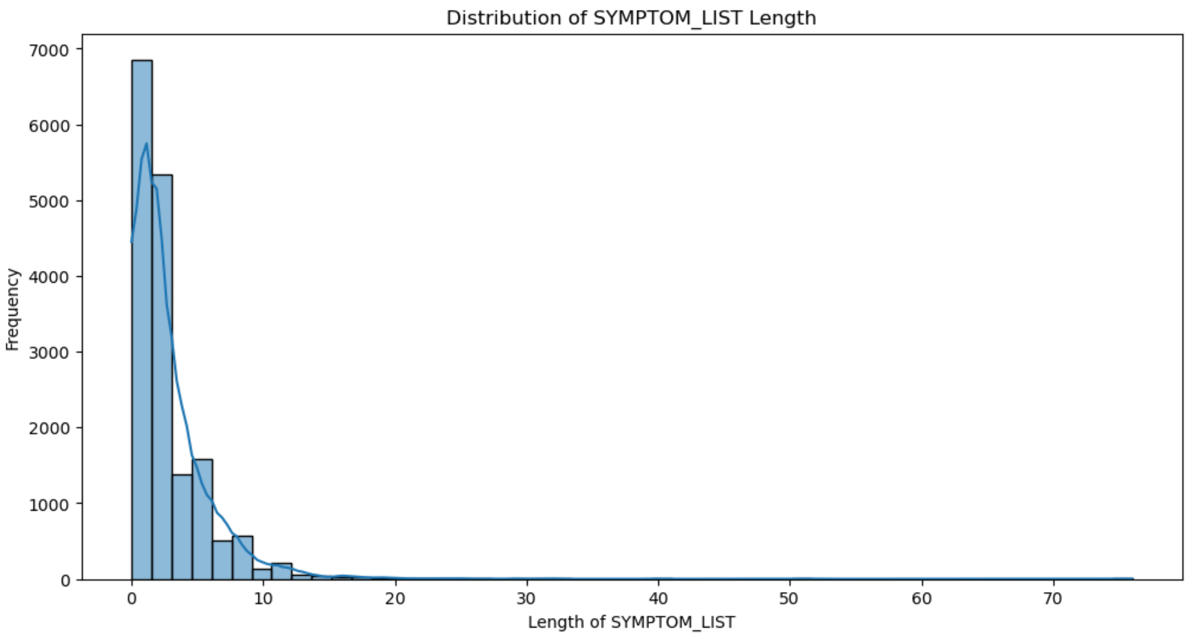


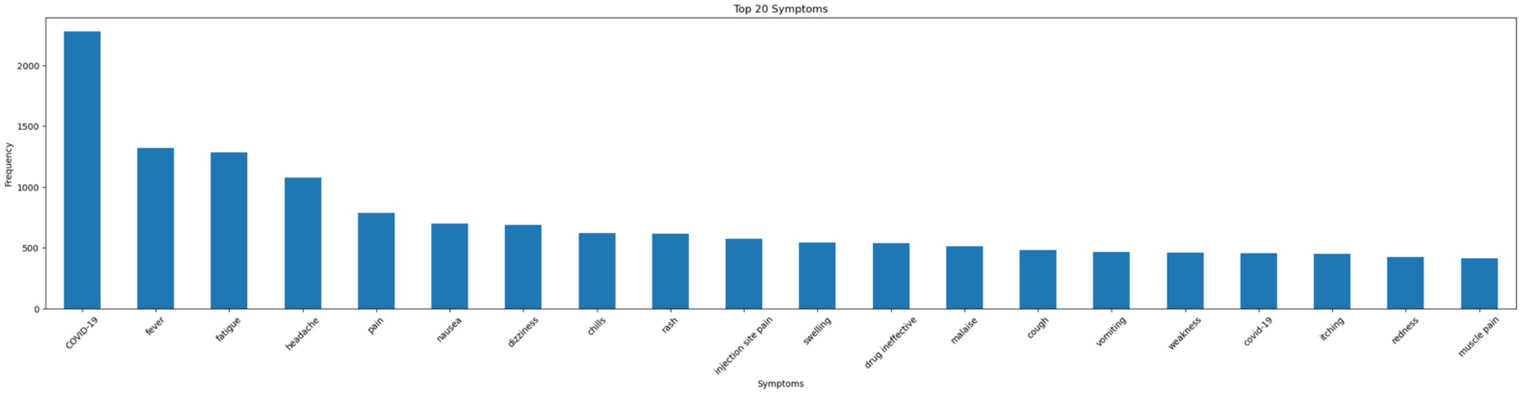




- Symptoms were extracted using LLM (Gemini-1.5-Flash) without any pre-processing and prompted to maintain a symptom dictionary to avoid synonyms.







Temporal Analysis

- Temporal relationship analysis in Natural Language Processing (NLP) involves identifying and understanding the temporal (time-based) relationships between entities
- It is a critical task in many domains, such as medicine, legal texts, and news, where timelines and event relationships are essential for decision-making.
- **Why Temporal Relationship Analysis for Symptom Extraction?**
- **Understanding Onset Patterns:**
  - Temporal analysis helps determine when symptoms began, which is critical for assessing causality or side effects.
- Progression Tracking:
  - By analyzing temporal data, researchers can trace symptom development and severity.
- Resolution and Outcome:
  - Temporal analysis can identify when symptoms resolve or persist, offering insights into long-term effects.

Metrics:

- **1. Kendall's Tau Rank Correlation Coefficient**
  - A statistical measure that evaluates the strength of the relationship between two ranked lists (e.g., the order of symptoms in two sequences).
  - It checks whether the relative order of symptoms in one sequence matches the order in the other.
    - Counts **concordant pairs**: Pairs of symptoms that have the same order in both sequences.
    - Counts **discordant pairs**: Pairs of symptoms where the order is reversed between the sequences.
    - Values range from -1 (completely reversed order) to 1 (perfectly aligned order), with 0 indicating no correlation.

- 2. Longest Common Subsequence (LCS)
    - A sequence comparison metric that identifies the longest subsequence of symptoms appearing in the same order in two sequences.
    - Unlike exact matches, LCS allows skipping symptoms that don't match while preserving the relative order.
    - Evaluates partial matches and tolerates differences in symptom sequences, making it ideal for analyzing noisy or incomplete data.
    - Provides insight into how closely symptom progressions align in structure.

- 3. Dynamic Time Warping (DTW)

A similarity measure designed for sequences that may differ in timing or speed, such as symptom progressions that vary in duration across patients.

- Handles variations in symptom timing across patients (e.g., one patient develops a fever earlier than another).
- Ideal for evaluating sequences where the order is important, but the timing of symptoms can vary.

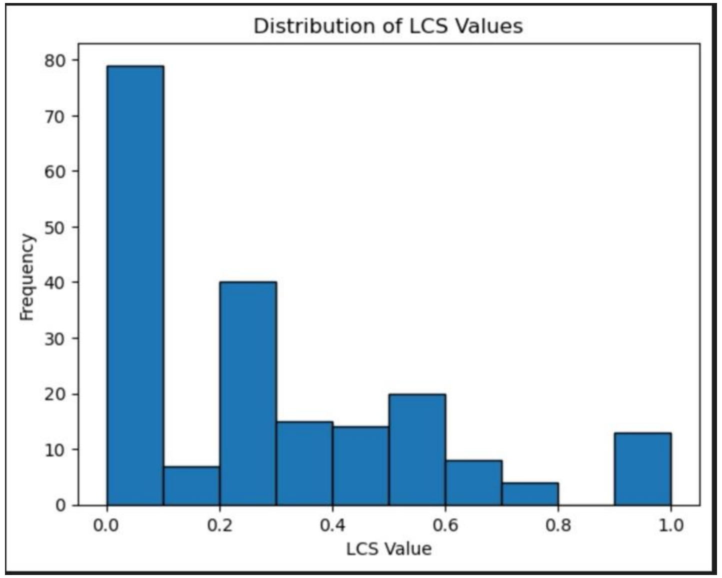
Results

- For Kendal Tau:
  - Aggregated value for 200 datapoints:(NER) **0.1290031915542915**
  - Aggregated value for 200 datapoints:(LLM) **0.33926521910912494**

```
[-0.7071067811865477,
 0.18257418583505539,
 0.31622776601683794,
 -0.9128709291752769,
 0.35856858280031806,
 0.33333333333333337,
 0.31622776601683794,
 0.0,
 0.0,
 -0.2357022603955159,
 -0.08606629658238703,
 0.5976143046671968,
 0.5976143046671968,
 0.0,
 0.7378647873726218,
 0.7071067811865477,
 0.7071067811865477,
 0.6324555320336759,
 -0.31622776601683794,
 0.5976143046671968,
 -0.6324555320336759,
 -0.18257418583505539,
 0.447213595499958,
 0.0,
 0.0,
...
 1.0,
 1.0,
 1.0,
 1.0,
 -1.0]
```

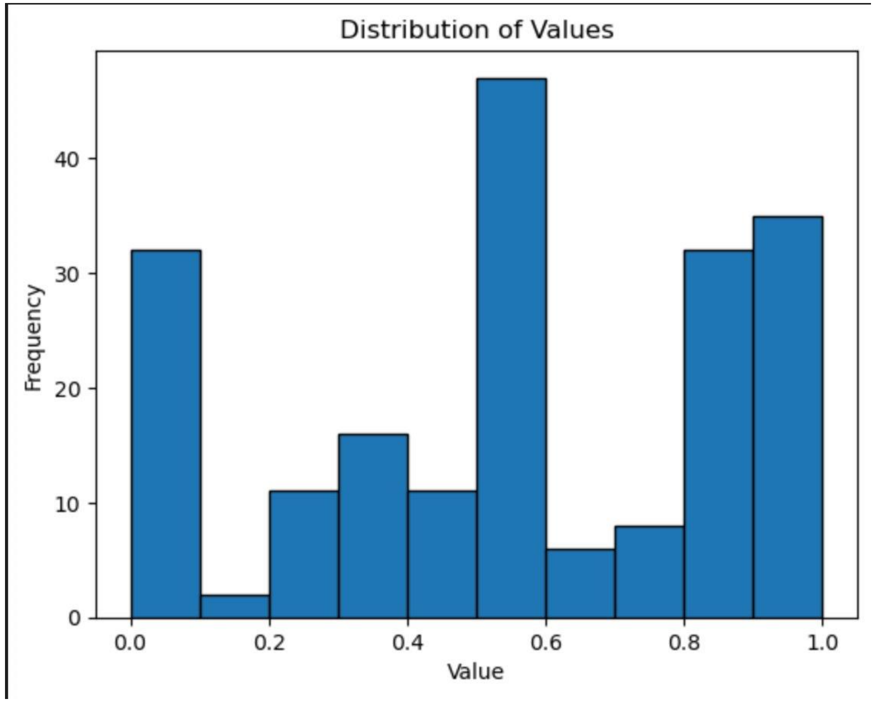
```
[0.3333333333333334,
 0.3333333333333334,
 0.39999999999999997,
 -0.5477225575051662,
 0.35856858280031806,
 -0.6666666666666669,
 0.11952286093343936,
 0.6,
 0.6666666666666669,
 1.0,
 -0.2,
 0.19999999999999998,
 -0.19999999999999998,
 0.39999999999999997,
 0.19999999999999998,
 0.39999999999999997,
 0.7999999999999999,
 0.9999999999999999,
 0.9999999999999999,
 0.6,
 0.7999999999999999,
 0.0,
 -0.2,
 0.7999999999999999,
 0.39999999999999997,
...
 1.0,
 1.0,
 1.0,
 1.0,
 1.0]
```

- LCS VALUES:
- NER 200 points distribution:



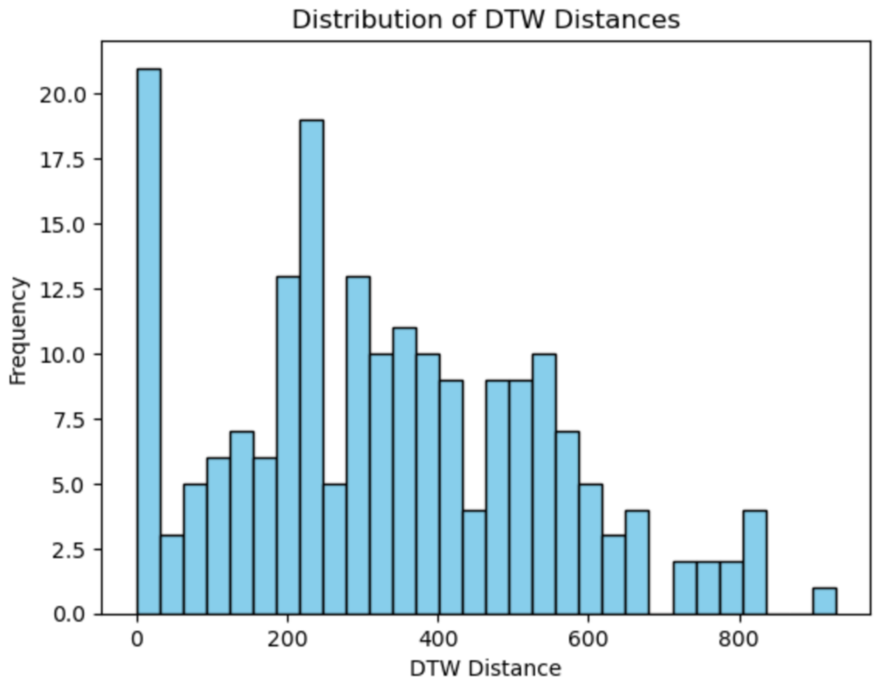

```
[0.0,
 0.0,
 0.0,
 0.0,
 0.0,
 0.3333333333333333,
 0.4,
 0.0,
 0.25,
 0.0,
 0.16666666666666666,
 0.4,
 0.2,
 0.0,
 0.2,
 0.0,
 0.0,
 0.0,
 0.0,
 0.2,
 0.0,
 0.25,
 0.16666666666666666,
 0.0,
 0.2,
...
 0.5,
 1.0,
 0.5,
 0.5,
 0.0]
```

- LLM 200 points distribution:

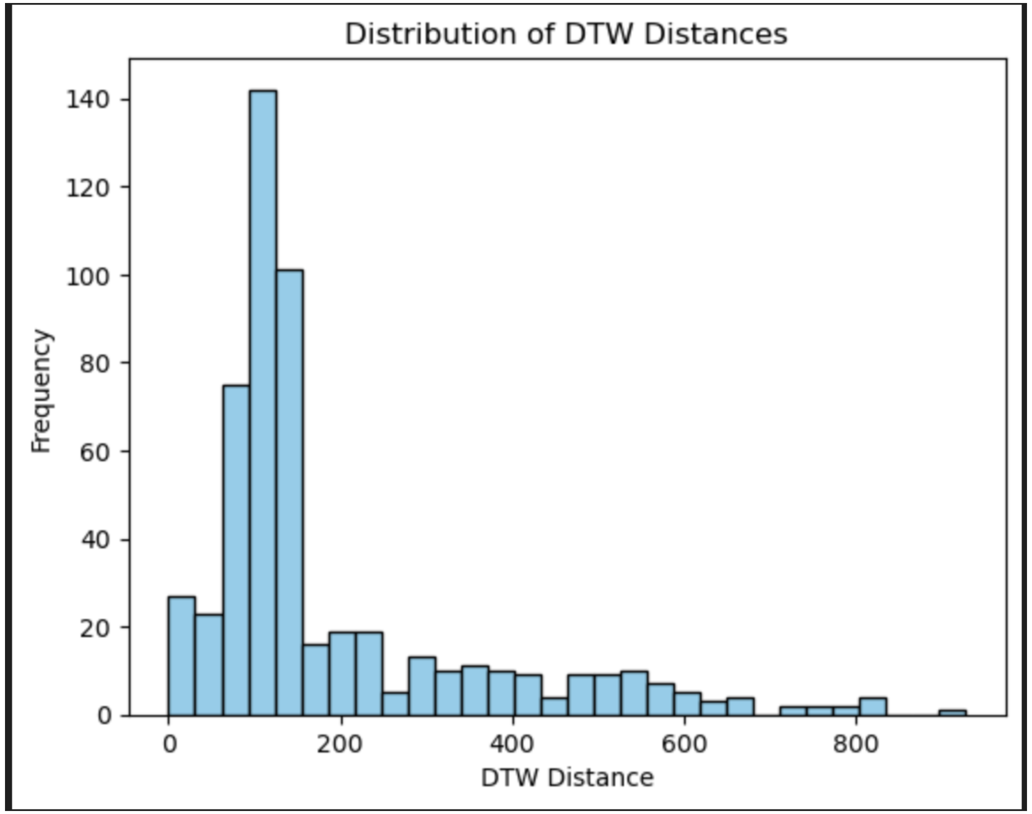



```
[0.75,
 0.5,
 0.8,
 0.25,
 0.0,
 0.5,
 0.0,
 0.8,
 0.5,
 0.5,
 0.3333333333333333,
 0.8,
 0.6,
 0.8,
 0.8,
 0.8,
 0.8,
 1.0,
 1.0,
 0.8,
 0.8,
 0.8,
 0.8333333333333334,
 0.8,
 0.6,
...
 1.0,
 1.0,
 0.5,
 0.5,
 0.5]
```

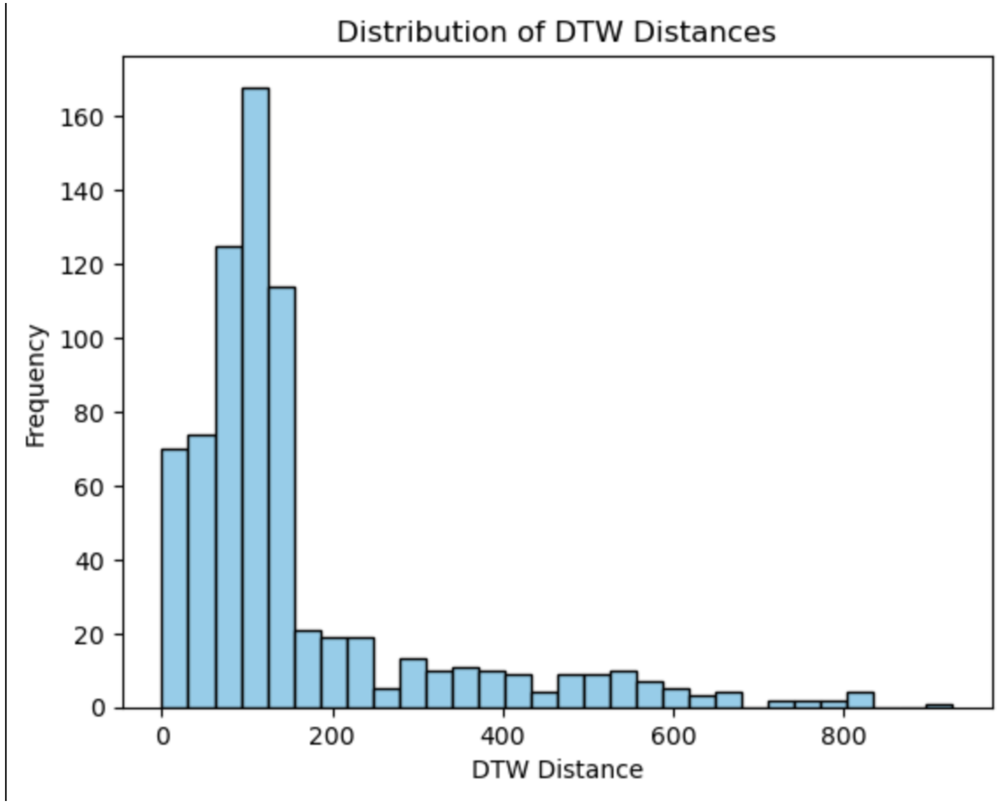
- DTW distribution for 200 datapoints:
  - This is for NER. We realized it is distance of words so we thought rather than finding distances between number, find distances between word embeddings for better result. Used Word2Vec.



- DTW for 200 data points: NER



- LLM



Associative Analysis

Apriori Algorithm

- **Associative Rule Mining – Market Basket Analysis / Biomedical Applications**
- Identifies frequent patterns in datasets by analyzing item co-occurrence.
- **Detects symptom co-occurrences** from VAERS reports. Helps identify associations like:
- Fever ⇔ Headache
- Dizziness ⇔ Fatigue
- Provides insights into **symptom progression**

How It Works

Identify frequent itemsets

(symptoms that occur together frequently).

Generate association rules

(e.g., "If Fever, then Headache with 80% confidence").

Evaluate using metrics

- **Support:** How often symptoms occur together.
- **Confidence:** Likelihood of symptom B given symptom A.
- **Lift:** Strength of the association compared to random chance.

$$Rule: \; X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

VAERS Symptoms (15k+)

Top 20 frequent symptoms.

```
- Injection site pain
- Loss of personal independence in daily activities
- Pain in extremity
- X-ray
- Angiogram
- Angiogram cerebral
- Aphasia
- Blood glucose
- CSF cell count
- Alpha haemolytic streptococcal infection
- Blood culture positive
- Chills
- Computerised tomogram thorax abnormal
- Endocarditis
- Decreased appetite
- Diarrhoea
- Fatigue
- Night sweats
- Arthralgia
- Urticaria
```

Top 20 frequent itemsets.

	support	itemsets
7	0.163670	(COVID-19)
25	0.101547	(Expired product administered)
16	0.089097	(No adverse event)
35	0.071366	(Drug ineffective)
4	0.070674	(Fatigue)
60	0.067027	(Drug ineffective, COVID-19)
38	0.055835	(Headache)
32	0.051496	(SARS-CoV-2 test)
33	0.049862	(Vaccination failure)
59	0.049296	(COVID-19, Vaccination failure)
58	0.045775	(COVID-19, SARS-CoV-2 test)
5	0.045586	(Arthralgia)
18	0.044014	(Dizziness)
2	0.036532	(Chills)
0	0.035777	(Injection site pain)
41	0.035651	(Pyrexia)
26	0.035337	(Pain)
8	0.035086	(Asthenia)
1	0.033828	(Pain in extremity)
31	0.031627	(Product administered to patient of inappropri...

VAERS Symptoms (15k+)

Association Rules.

	antecedents	consequents	support	confidence	lift
7	(Injection site erythema)	(Injection site swelling)	0.014462	0.516854	23.286246
6	(Injection site swelling)	(Injection site erythema)	0.014462	0.651558	23.286246
8	(Underdose)	(Product administered to patient of inappropri...	0.010501	0.609489	19.271002
9	(Headache, Malaise)	(Fatigue)	0.010123	0.797030	11.277545
10	(Fatigue, Malaise)	(Headache)	0.010123	0.624031	11.176339
2	(Body temperature)	(Fatigue)	0.011255	0.617241	8.733636
1	(Myalgia)	(Fatigue)	0.015279	0.532895	7.540176
0	(Malaise)	(Fatigue)	0.016222	0.522267	7.389802
4	(Vaccination failure)	(COVID-19)	0.049296	0.988651	6.040530
12	(Drug ineffective, SARS-CoV-2 test)	(COVID-19)	0.019995	0.984520	6.015293
11	(Vaccination failure, SARS-CoV-2 test)	(COVID-19)	0.019429	0.984076	6.012582
5	(Drug ineffective)	(COVID-19)	0.067027	0.939207	5.738436
3	(SARS-CoV-2 test)	(COVID-19)	0.045775	0.888889	5.430998

NLP Method – Data Preprocessing + NER from VAERS Data

Top 20 frequent symptoms.

```
- sore
- muscle pain
- arexvy
- coronavirus disease 2019
- pain
- seizures
- arrest
- seizure
- auto-immune disease
- encephalitis
- endocarditis
- fever
- chills
- infection
- pains fatigue chills
- diarrhea loss
- fatigue diarhea loss
- prolonged shoulder pain
- nan
- respiratory failure s/p trach placement 8/17/2023 gerd hypertension
```

Top 20 frequent itemsets.

	support	itemsets
8	0.263581	(coronavirus disease)
1	0.145687	(pain)
2	0.078094	(coronavirus disease 2019)
64	0.063443	(coronavirus disease, coronavirus disease 2019)
32	0.060991	(swelling)
5	0.058224	(fever)
15	0.056527	(fatigue)
21	0.047598	(headache)
54	0.046403	(pain, coronavirus disease)
26	0.039424	(allergies)
59	0.035274	(pain, swelling)
77	0.033828	(coronavirus disease, allergies)
33	0.033262	(rash)
55	0.031439	(pain, fatigue)
24	0.029049	(allergy)
3	0.026031	(sore)
73	0.025402	(coronavirus disease, fatigue)
36	0.025025	(redness)
56	0.024962	(pain, headache)
65	0.024899	(fever, coronavirus disease)

NLP Method – Data Preprocessing + NER from VAERS Data

Association Rules (647 rows with lift > 1.0) [Top 40]

	antecedents	consequents	support	confidence	lift
470	(pain, pyrexia)	(fever, erythema)	0.010312	0.689076	61.567746
473	(fever, erythema)	(pain, pyrexia)	0.010312	0.921348	61.567746
281	(pain, pyrexia)	(fever, muscle pain)	0.010123	0.676471	55.172247
283	(fever, muscle pain)	(pain, pyrexia)	0.010123	0.825641	55.172247
635	(pain, myalgia, fatigue)	(erythema, muscle pain)	0.010501	0.625468	51.809613
639	(erythema, muscle pain)	(pain, myalgia, fatigue)	0.010501	0.869792	51.809613
622	(swelling, myalgia)	(pain, muscle pain, fatigue)	0.011004	0.911458	51.770833
605	(pain, muscle pain, fatigue)	(swelling, myalgia)	0.011004	0.625000	51.770833
628	(pain, muscle pain, fatigue)	(erythema, myalgia)	0.010501	0.596429	51.552174
645	(erythema, myalgia)	(pain, muscle pain, fatigue)	0.010501	0.907609	51.552174
418	(swelling, myalgia)	(muscle pain, fatigue)	0.011129	0.921875	50.731834
423	(muscle pain, fatigue)	(swelling, myalgia)	0.011129	0.612457	50.731834
615	(muscle pain, fatigue)	(pain, swelling, myalgia)	0.011004	0.605536	50.686578
614	(pain, swelling, myalgia)	(muscle pain, fatigue)	0.011004	0.921053	50.686578
612	(pain, myalgia, fatigue)	(swelling, muscle pain)	0.011004	0.655431	50.601796
617	(swelling, muscle pain)	(pain, myalgia, fatigue)	0.011004	0.849515	50.601796
433	(muscle pain, fatigue)	(erythema, myalgia)	0.010563	0.581315	50.245825
428	(erythema, myalgia)	(muscle pain, fatigue)	0.010563	0.913043	50.245825
637	(pain, erythema, myalgia)	(muscle pain, fatigue)	0.010501	0.912568	50.219676
638	(muscle pain, fatigue)	(pain, erythema, myalgia)	0.010501	0.577855	50.219676

	antecedents	consequents	support	confidence	lift
642	(myalgia, fatigue)	(pain, erythema, muscle pain)	0.010501	0.596429	49.924211
631	(pain, erythema, muscle pain)	(myalgia, fatigue)	0.010501	0.878947	49.924211
589	(pain, myalgia, fatigue)	(muscle pain, headache)	0.011004	0.655431	49.875455
594	(muscle pain, headache)	(pain, myalgia, fatigue)	0.011004	0.837321	49.875455
544	(pain, myalgia, fatigue)	(fever, muscle pain)	0.010249	0.610487	49.790685
548	(fever, muscle pain)	(pain, myalgia, fatigue)	0.010249	0.835897	49.790685
430	(fatigue, myalgia)	(erythema, muscle pain)	0.010563	0.600000	49.700000
431	(erythema, muscle pain)	(fatigue, myalgia)	0.010563	0.875000	49.700000
608	(pain, swelling, muscle pain)	(myalgia, fatigue)	0.011004	0.870647	49.452736
620	(myalgia, fatigue)	(pain, swelling, muscle pain)	0.011004	0.625000	49.452736
547	(muscle pain, fatigue)	(pain, fever, myalgia)	0.010249	0.564014	49.286133
546	(pain, fever, myalgia)	(muscle pain, fatigue)	0.010249	0.895604	49.286133
585	(pain, muscle pain, headache)	(myalgia, fatigue)	0.011004	0.866337	49.207921
597	(myalgia, fatigue)	(pain, muscle pain, headache)	0.011004	0.625000	49.207921
540	(pain, fever, muscle pain)	(myalgia, fatigue)	0.010249	0.862434	48.986243
551	(myalgia, fatigue)	(pain, fever, muscle pain)	0.010249	0.582143	48.986243
421	(swelling, muscle pain)	(fatigue, myalgia)	0.011129	0.859223	48.803883
420	(fatigue, myalgia)	(swelling, muscle pain)	0.011129	0.632143	48.803883
553	(fever, myalgia)	(pain, muscle pain, fatigue)	0.010249	0.853403	48.473298
537	(pain, muscle pain, fatigue)	(fever, myalgia)	0.010249	0.582143	48.473298

LLM Method – Gemini 1.5 Flash from VAERS Data

Top 20 frequent symptoms.

```
- muscle pain
- arm soreness
- pain
- activities of daily living impaired
- micro-seizures
- seizures
- arrest in speech
- eye rolling
- arm stiffness
- seizure
- encephalitis
- endocarditis
- fever
- chills
- lung infection
- aches and pains
- fatigue
- night sweats
- diarrhea
- loss of appetite
```

Top 20 frequent itemsets.

	support	itemsets
8	0.135548	(COVID-19)
2	0.078381	(fever)
4	0.076242	(fatigue)
12	0.063882	(headache)
0	0.045460	(pain)
13	0.041478	(nausea)
17	0.040944	(dizziness)
3	0.036784	(chills)
35	0.036665	(rash)
29	0.034050	(injection site pain)
32	0.031852	(drug ineffective)
14	0.031792	(swelling)
19	0.030425	(malaise)
20	0.028464	(cough)
21	0.027633	(vomiting)
10	0.027217	(weakness)
45	0.026919	(covid-19)
37	0.025790	(itching)
64	0.025671	(headache, fatigue)
15	0.025256	(redness)

LLM Method – Gemini 1.5 Flash from VAERS Data

Association Rules (62 rows with lift > 1.0) [Top 40]

	antecedents	consequents	support	confidence	lift
61	(injection site swelling)	(injection site redness, injection site pain)	0.011469	0.654237	47.660195
58	(injection site redness, injection site pain)	(injection site swelling)	0.011469	0.835498	47.660195
60	(injection site redness)	(injection site pain, injection site swelling)	0.011469	0.696751	46.712845
59	(injection site pain, injection site swelling)	(injection site redness)	0.011469	0.768924	46.712845
20	(injection site redness)	(injection site swelling)	0.012836	0.779783	44.482017
21	(injection site swelling)	(injection site redness)	0.012836	0.732203	44.482017
54	(injection site pain, fatigue)	(injection site tenderness)	0.012123	0.625767	38.857583
56	(injection site tenderness)	(injection site pain, fatigue)	0.012123	0.752768	38.857583
52	(injection site redness)	(injection site pain, fatigue)	0.010459	0.635379	32.798033
51	(injection site pain, fatigue)	(injection site redness)	0.010459	0.539877	32.798033
48	(injection site swelling)	(injection site pain, fatigue)	0.011112	0.633898	32.721597
46	(injection site pain, fatigue)	(injection site swelling)	0.011112	0.573620	32.721597
31	(injection site pain, malaise)	(muscle pain)	0.010340	0.731092	29.716965
47	(injection site swelling, fatigue)	(injection site pain)	0.011112	0.963918	28.308559
50	(injection site redness, fatigue)	(injection site pain)	0.010459	0.956522	28.091357
55	(injection site tenderness, fatigue)	(injection site pain)	0.012123	0.948837	27.865676
30	(muscle pain)	(injection site pain, fatigue)	0.013073	0.531401	27.430722
28	(injection site pain, fatigue)	(muscle pain)	0.013073	0.674847	27.430722
4	(injection site tenderness)	(muscle pain)	0.010696	0.664207	26.998235
33	(malaise, muscle pain)	(injection site pain)	0.010340	0.915789	26.895123

	antecedents	consequents	support	confidence	lift
19	(injection site tenderness)	(injection site pain)	0.014678	0.911439	26.767360
57	(injection site redness, injection site swelling)	(injection site pain)	0.011469	0.893519	26.241064
25	(malaise, fatigue)	(muscle pain)	0.010459	0.637681	25.920045
17	(injection site swelling)	(injection site pain)	0.014916	0.850847	24.987890
18	(injection site redness)	(injection site pain)	0.013727	0.833935	24.491202
29	(muscle pain, fatigue)	(injection site pain)	0.013073	0.794224	23.324954
32	(injection site pain, muscle pain)	(malaise)	0.010340	0.696000	22.875563
44	(malaise, fatigue)	(injection site pain)	0.012657	0.771739	22.664618
43	(injection site pain, fatigue)	(malaise)	0.012657	0.653374	21.474573
26	(muscle pain, fatigue)	(malaise)	0.010459	0.635379	20.883123
2	(joint pain)	(muscle pain)	0.011944	0.505025	20.527930
16	(redness)	(swelling)	0.015391	0.609412	19.168563
3	(muscle pain)	(injection site pain)	0.014856	0.603865	17.734443
15	(vomiting)	(nausea)	0.017174	0.621505	14.983800
24	(malaise, muscle pain)	(fatigue)	0.010459	0.926316	12.149682
42	(injection site pain, malaise)	(fatigue)	0.012657	0.894958	11.738389
27	(injection site pain, muscle pain)	(fatigue)	0.013073	0.880000	11.542198
40	(injection site pain, headache)	(fatigue)	0.011766	0.860870	11.291281
53	(injection site pain, injection site tenderness)	(fatigue)	0.012123	0.825911	10.832758
36	(injection site pain, fever)	(fatigue)	0.010043	0.820388	10.760324

Summary

- VAERS contains valuable but complex data.
- NLP & LLMs uncover insights from unstructured text.
- Hybrid models outperform traditional approaches

Approach	Advantages	Challenges
NER	Faster, lower cost	Misses' semantic nuances
LLM (GPT-4o)	Handles complex semantics	Expensive, hard to scale
LLM (Gemini 1.5 Flash)	Handles complex semantics	Comparatively cheaper than 4o, longer context handling and faster output speed

THANK YOU!

Any questions?