Predictive Modeling and Statistical Analysis of Serious Adverse Events Following Shingles Vaccination Using VAERS Data

Author: Youngmin Choi, M.S. Candidate,

Health Informatics & Data Science, Georgetown University

Industry Sponsor/Mentor: Dr. Azade Tabaie, Data Scientist, MedStar Health

Faculty Academic Mentor: Dr. Peter McGarvey, Director, ICBI, Georgetown University

Executive Summary

Understanding the factors influencing serious adverse events (SAEs) following vaccination is crucial among elderly population for informed decision-making. This study utilizes the Vaccine Adverse Event Reporting System (VAERS) dataset to develop a predictive model for SAE occurrence following the onset of adverse events (AEs) after immunization. Using natural language processing (NLP) techniques and machine learning algorithms, we developed an AI-based predictive model for SAE and identified key predictors of SAE, including specific medical conditions, age, and gender. While this study demonstrates the potential of data-driven approaches for vaccine safety monitoring among elderly population, it also indicates challenges associated with VAERS dataset, and future research should aim to validate the findings in this study. By providing insights into SAE occurrence after the onset of AEs, the purpose of this study is to contribute to informative and personalized decisions, ultimately to enhance safety and compliance of immunization.

Introduction

Shingles (herpes zoster) is a painful skin rash that develops on one side of the face or body, caused by varicella-zoster virus (VZV), which remains in the body after chickenpox.[1] Centers for Disease Control and Prevention (CDC) recommends that adults 50 years and older get shingles (HZ) vaccine injections, to prevent shingles and related complications. Shingrix is a recombinant shingles vaccine developed by GlaxoSmithKline for shingles prevention. It showed 97% effectiveness in preventing shingles in adults from 50 to 69 years old, and 91% effectiveness in adults 70 years old and older. While Shingrix is generally safe, studies have reported side effects including sore arm, redness and swelling at the site of injection, and Guillain-Barré syndrome (GBS), a serious condition, was also reported although rarely.[2] However, according to Wang et al. (2023), only 56.06% of eligible adults had the willingness to receive HZ vaccination. Main reasons for such hesitancy included concerns about the safety of the vaccine among other factors.[3] The World Health Organization (WHO) identified vaccine hesitancy as one of its top threats to global health and increase in vaccine confidence as its solution.[4] In order to increase vaccine confidence, it is important to promote accurate information. A data-driven approach to vaccine safety information is crucial for increasing vaccine confidence by providing evidence-based facts. This will allow healthcare providers, elderly population, and their care givers to make informed and personalized decisions about vaccination.

AEs are any undesirable experience related to the medical product including vaccines in a patient, whereas SAEs are defined as events with serious patient outcomes including death, life-threatening, hospitalization and/or prolonged hospitalization, or disability/permanent damage.[5] Like other medical conditions or disease, early detection of such symptom or sign of progress is important for early intervention, in order to mitigate disease progression and better patient outcome.[6] VAERS is a database created by the Food and Drug Administration (FDA) and CDC to monitor safety data of vaccines and detect any AEs not identified in pre-market clinical trials. Although VAERS does collect comprehensive information, it is a passive surveillance system with unverified reports and has limitations such as subjective, biased, and/or under-reporting.

In this study, the objective is to predict the progression of AE following shingles vaccination into SAE by developing an AI-based predictive model using the data from the VAERS database with Python, based on symptom description, patients' demographics (i.e., gender, age) and current illness as well as medical history information. The model utilizes Natural Language Processing (NLP) tools to process symptom description, current illness, and medical history information to create vector embeddings of this information. Term Frequency-Inverse Document Frequency (TF-IDF) and Bidirectional Encoders from Transformers (BERT) were used as text processing tools. For machine learning models, Logistic Regression (LR), LR with elastic net regularization (EN), and Extreme Gradient Boosting (XGBoost) were studied in this project. Feature importance analysis was performed to identify key phrases related to medical terminology influencing predictions. In addition, statistical analyses were conducted to assess any indications of correlations between certain pre-existing medical conditions, age group, or gender with prevalence of SAEs.

By developing this model and conducting statistical analysis, this project aims to provide valuable insights into shingles vaccine safety, potentially increasing vaccination confidence and improving elderly population health outcomes.

Methods and Analysis

Data Source

VAERS is a publicly available dataset, which was downloaded from the VAERS website. Shingrix, the vaccine of interest in this study, was approved by the FDA in October 2017, and therefore data from year 2018 to 2023 were processed in this study. The dataset is comprised of 3 separate comma separated value (CSV) files for each year; 'VAERS DATA', 'VAERS Symptoms', and 'VAERS Vaccine'. the 'VAERS DATA' file contains comprehensive information of each of the reports, including patient information, symptom description, date of report, current illness, emergency room visit, hospitalization, death, and so on. 'VAERS Symptoms' file contains adverse event terms from MedDRA dictionary, which were selected by designated coders. 'VAERS Vaccine' provides information related to vaccines that were administered including manufacturer, product name, and lot number. These tables share a common column which is the 'VAERS_ID' column, which is a uniquely designated number for each report. In this study, free-text entries from the symptom description, current illness, and history columns, alongside demographic information such as age and gender from the 'VAERS DATA' table, were processed for predictive modelling. Vaccine related information was sourced from the 'VAERS Vaccine' table. These tables were processed further for predictive model development and subsequent statistical analysis.

Data Cleaning

VAERS DATA tables from year 2018 to 2023 were merged into one table with 50,267 rows and were filtered to prepare the elderly population data for model development and analysis with 2 criteria: i) age of patients 50 years and older, ii) vaccine name 'ZOSTER (SHINGRIX). This filtration process was conducted to ensure that only the data related to vaccine of interest, as used in the intended population, was included in this study.

Furthermore, 7 rows of data which had missing values in the 'SYMPTOM_TEXT' column were removed for data completeness. In addition, 41 rows which described patients' symptoms as 'None stated'. These rows were removed as they did not provide any meaningful information for the model and may introduce noise to the model. Rows with unknown gender ('U') were also filtered out, resulting in a final dataset of 49,669 entries after cleaning. Among these, 45,084 entries were classified as the non-SAE group and 4,585 as the SAE group.

Text Preprocessing of Symptom Text

For contextual word embedding of free-text based columns, 2 distinct tools were utilized; TF-IDF and BERT. Contextual word embedding, converts text data into numeric vectors, which are then used for subsequent model training.[7]

Different text preprocessing steps were applied for TF-IDF and BERT, respectively. This is due to the distinct approaches each tool uses for embedding calculations. While TF-IDF calculates the embedding features based on the frequency, necessitating preprocessing steps such as removing certain stop words and applying stemming. These preprocessing steps optimize TF-IDF to capture significant word occurrences while minimizing noise. In contrast, BERT derives them by considering the context of each word from both directions (left to right, right to left) within the sentence. Therefore, minimal text preprocessing steps were employed on BERT preparation in order to keep the context intact as the original text. For TF-IDF embedding preparation, date and time texts (e.g., 10:00) were substituted with designated terms. Punctuations were removed, and all texts were converted to lowercase letters. A pre-defined list of stop words were removed as well and for the last step, stemming was applied to complete the text preprocessing. For the preparation of BERT text embeddings, date and time text substitution, lowercase transformation and punctuation removal steps were incorporated.[8, 9] Additionally, words including 'shingrix' were excluded from BERT preprocessing to keep focus on general patient symptom expressions and medical contexts, in comparison to vaccine-specific details. Similarly, one-letter words were also removed to reduce irrelevant noise in the model input. These decisions aim to enhance the relevance and effectiveness of BERT embeddings in capturing meaningful contextual information from the text data.

Text Preprocessing of Current Illness & History

In order to incorporate patients' pre-existing medical condition, both 'CUR_ILL' and 'HISTORY' columns were utilized in this study. Texts in these columns were converted to lowercase, and certain non-informative words were removed, such as 'none', 'unknown', 'no other problems', and 'no known'. These words indicated that the patient had no pre-existing health conditions, therefore removing them in this step helped eliminate noise from the dataset. After cleaning the text as described, both columns were concatenated along with the preprocessed symptom text. This approach was necessary since more than half of the dataset had missing values in both the 'CUR_ILL' and 'HISTORY' columns. By concatenating these columns, the embedding could capture the contextual information and handle the rows with missing data more effectively.

Data Preprocessing

Following text data processing, the dataset was split into train and test sets with 80 to 20 ratio. Due to the high imbalance in the dataset, stratified split was conducted.

Text Embedding

TF-IDF and BERT were used to generate embeddings of texts from columns 'SYMPTOM_TEXT', 'CUR_ILL', and 'HISTORY'. TF-IDF is a technique where it evaluates the words used in a certain text and compares to words used in the whole document, which means it captures each words' relevance in the entire document. Words with high TF-IDF values indicate a strong relationship with the document.[8] BERT is an NLP model developed by Google, which is able to identify the context of a certain word by understanding its context bidirectionally and produces contextual embeddings. This allows BERT to capture nuanced meanings and relationships between words, resulting in more accurate text representations. This study also utilized BlueBERT, a BERT model pre-trained on PubMed abstracts and MIMIC-III clinical notes by the National Center for Biotechnology Information (NCBI). The purpose was to identify a better-performing embedding tool and to compare the performance with base BERT model. [10] The resulting embeddings were used as features to the classification models for SAE prediction.[9, 11]

Demographics Feature Preprocessing

'AGE_YRS' column, with numerical type, was included as one of the features. StandardScaler function in Python was used to scale the value to create a value range with a mean of 0 and standard deviation of 1. This scaling is important to ensure numerical values contribute appropriately to the model, which helps improving model performance by treating all features on the same level. [12]

The 'SEX' column, originally a categorical feature with 'M' and 'F' values, was one-hot encoded.

Imbalance Adjustment

Prior to model training, the highly imbalanced dataset was adjusted by random undersampling of the training set. The original data contained only approximately 10% SAE-positive cases. Undersampling the majority class was applied to the SAE-negative cases to achieve a final ratio of 7:3 between SAE-negative and SAE-positive cases. This ratio was chosen to maintain a representative distribution of the data while ensuring sufficient data for model training. While oversampling was also considered as an alternative, it was ultimately not applied due to the risk of overfitting. Given the model's need to train on numerous features which already increases the risk of overfitting, undersampling was selected as the method for imbalance adjustment.

Machine Learning Models

In this study, Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), and Elastic Net regression (EN) models were trained and tested for predictive modeling. LR is a statistical model used for binary classification, where it takes a set of input features and estimates the probability of the outcome belonging to a specific class. It has the advantage of implementing multiple variables, while reducing the influence of confounding factors.[13] XGBoost is a powerful decision tree-based boosting machine learning algorithm that combines the predictions of multiple decision trees, a process known as gradient boosting, to improve accuracy.[14] EN is a regularization method combining two regularization functions; L1 (Lasso) and L2 (Ridge). While EN can handle correlations between predictors with L2 regularization, sparsity is also obtained due to L1 regularization.[15] These three models were evaluated and their performances were compared.

Feature Importance

To understand important features, SHapley Additive exPlanations (SHAP) was used to extract features that were important in decision making process of the trained models. SHAP is a game theoretic approach to explain the output of machine learning models.[16] Following the SHAP analysis, the resulting list of important features were further refined. Words which were medically insignificant or lacking clinical relevance were removed to focus on words with potential prediction value. This step was crucial to ensure that subsequent interpretations were based on medically meaningful information.

Overall Modelling Process Description

Figure 1 describes the overall process flow of this study.

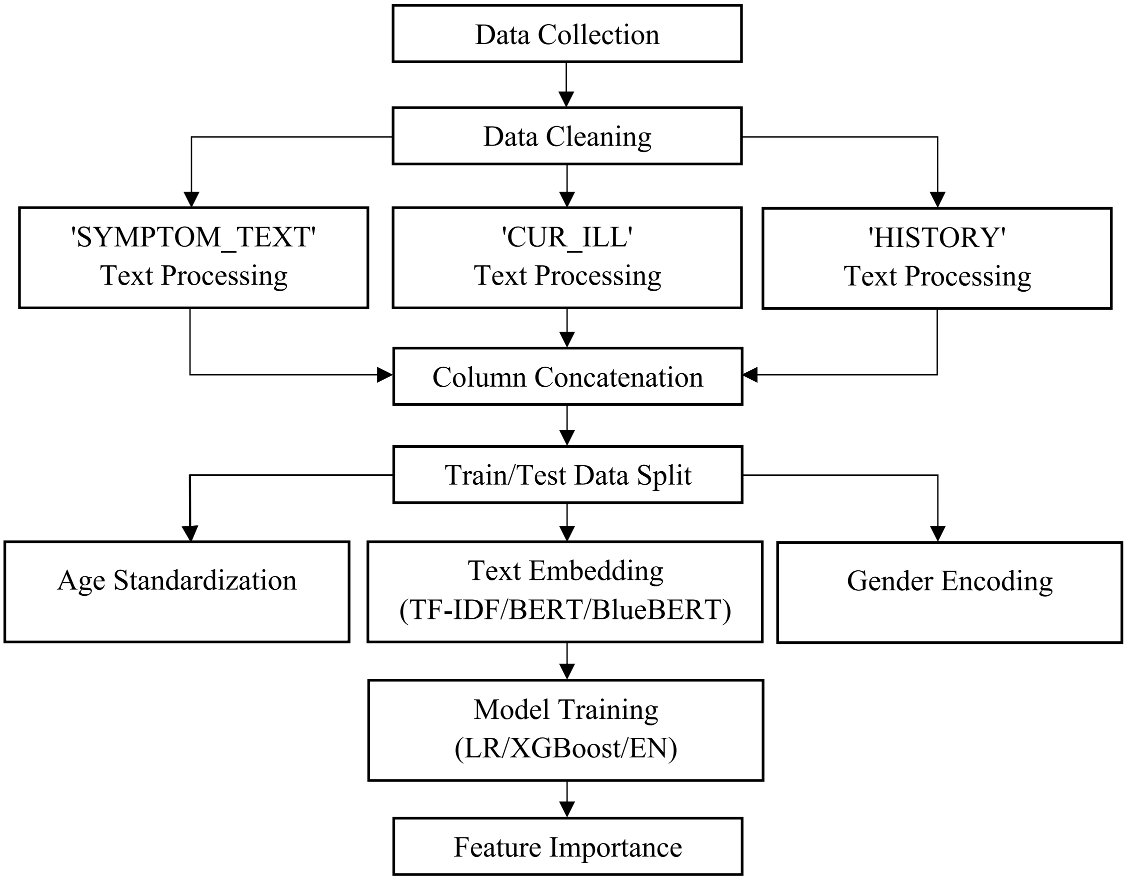
Data Collection
Data Cleaning
'SYMPTOM_TEXT' Text Processing
'CUR_ILL' Text Processing
'HISTORY' Text Processing
Column Concatenation
Train/Test Data Split
Age Standardization
Text Embedding (TF-IDF/BERT/BlueBERT)
Gender Encoding
Model Training (LR/XGBoost/EN)
Feature Importance

Fig. 1. Preprocessing and Modeling Flowchart. This flowchart describes the sequential steps that were taken to prepare the dataset for analysis, including data preprocessing, model training, and feature importance assessment.

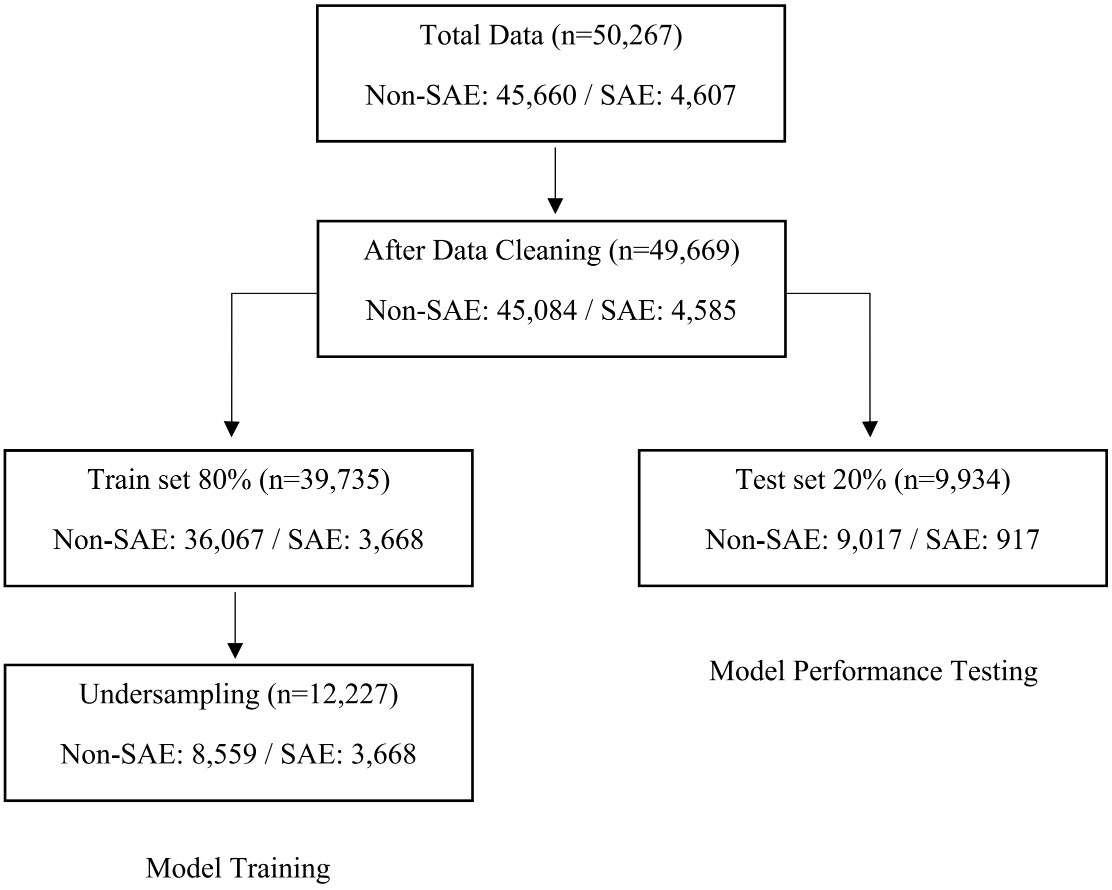
Total Data (n=50,267)
Non-SAE: 45,660 / SAE: 4,607
After Data Cleaning (n=49,669)
Non-SAE: 45,084 / SAE: 4,585
Train set 80% (n=39,735)
Non-SAE: 36,067 / SAE: 3,668
Test set 20% (n=9,934)
Non-SAE: 9,017 / SAE: 917
Undersampling (n=12,227)
Non-SAE: 8,559 / SAE: 3,668
Model Performance Testing
Model Training

Fig. 2. Data preprocessing results. This figure illustrates the outcome of data preprocessing, where 12,277 reports were allocated for model training and 9,934 reports were reserved for testing the trained models.

Subgroup Statistical Analysis

To investigate potential correlations between specific subgroups and the prevalence of SAE in this dataset, and odds ratio statistical analysis was performed. To ensure the accuracy of the analysis, data lacking both the 'CUR_ILL' and 'HISTORY' values were excluded, as there is a possibility that pre-existing conditions of these rows may not have been documented. The remaining text data from both columns underwent preprocessing, including lowercase conversion and removal of specific words and phrases. After concatenation, the processed text was analyzed using ScispaCy, a tool designed for analyzing and extracting information from text data. ScispaCy leverages the spaCy library which was retrained using datasets relevant to biomedical texts such as MedMentions dataset. [17, 18] In this study, entities with the label 'DISEASE' were extracted to identify medical conditions of each patient. Table 1 describes the frequency counts for each disease entity extracted through this process. The extracted diseases were then mapped to International

Classification of Disease 10th Edition (ICD) codes using a predefined mapping dataset. These codes were subsequently categorized into groups for further analysis. ICD code groups with number of instances lower than 100, specifically O00-O9A (n=1), P00-P96 (n=0), Q00-Q99 (n=77), and V00-Y99 (n=7), were excluded to secure statistical validity.

ICD Code Group	Description	Instance Count
A00 - B99	Certain infections and parasitic diseases	1,343
C00 - D49	Neoplasms	1,250
D50 - D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanisms	400
E00 - E89	Endocrine, nutritional and metabolic diseases	6,538
F01 - F99	Mental, behavioral and neurodevelopmental disorders	1,937
G00 - G99	Diseases of the nervous system	2,288
H00 - H59	Diseases of the eye and adnexa	599
H60 - H95	Diseases of the ear and mastoid process	212
I00 - I99	Diseases of the circulatory system	4,324
J00 - J99	Diseases of the respiratory system	2,455
K00 - K95	Diseases of the digestive system	1,417
L00 - L99	Diseases of the skin and subcutaneous tissue	690
M00 - M99	Diseases of the muscoskeletal system and connective tissue	4,664
N00 - N99	Diseases of the genitourinary system	754
O00 - O9A	Pregnancy, childbirth, and the puerperium	1
P00 - P96	Certain conditions originating in the perinatal period	0
Q00 - Q99	Congenital malformations, deformations and chromosomal abnormailities	77
R00 - R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	2,387
S00 - T88	Injury, poisoning and certain other consequences of external causes	854
U00 - U85	Codes for special purposes	10
V00 - Y99	External causes of morbidity	7
Z00 - Z99	Factors influencing health status and contact with health services	331

Table 1. ICD code groups and data distribution. The ICD code groups were defined as illustrated in this table, based on the classification by ICD10Data.com.[19] These counts represent the frequency of each code in the entire dataset. It should be noted that some patients have multiple ICD codes, leading to potential overlaps where patients may appear in several categories.

As for the age group, the distribution of ages of the whole dataset was initially assessed to determine the range of age groups. Subsequently, the age group definition was finalized as described in the following table based on this assessment.

Age Group	SAE Positive Frequency	SAE Negative Frequency
50 - 54	562	6,368
55 - 59	667	7,433
60 - 64	858	9,486
65 - 69	813	8,799
70 - 74	738	6,796
75 - 79	553	3,897
80 and older	416	2,881

Table 2. Age groups and data distribution. This table describes the age group definitions and their distribution in the dataset.

In the analysis, the 'SEX' column, representing categorical values for male ('M') and female ('F'), was included along with the defined conditions and age groups. Logistic regression was used to calculate odds ratios, assessing the association between these groups and SAE prevalence. To account for the effects of different groups, a comprehensive analysis was conducted incorporating all groups simultaneously.

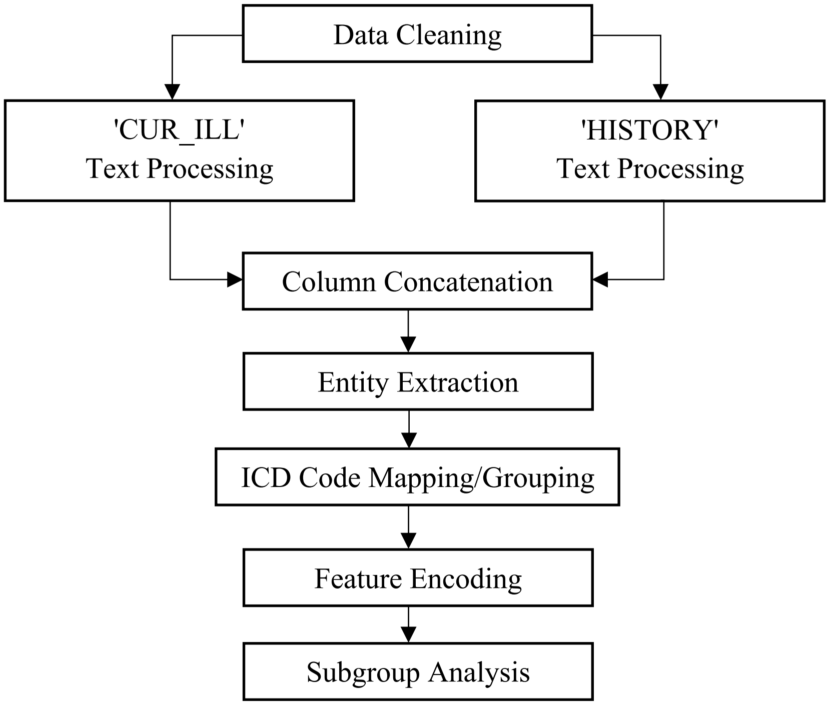

Fig. 3. Analysis Process. This figure illustrates the comprehensive process used for subgroup statistical analyses.

Results

For this study, a total of 50,267 reports related to patients who received the Shingrix vaccine from year 2018 to 2023 were included. Following data cleaning and filtration, a total of 49,669 reports were divided into training set and test set with a ratio of 80 to 20. The training set consisted of 39,735 reports, whereas the test set comprised 9,934 reports (Fig. 2). The mean age of the patients in this data was approximately 64.96 years old, with the maximum age recorded at 117 years.

Model Performance

The performance of the models was evaluated using several metrics, including area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F-1 score, area under the precision-recall curve (AUPRC), and accuracy. The combined texts of 'SYMPTOM_TEXT', 'CUR_ILL', and 'HISTORY' were processed using TF-IDF, BERT, and BluBERT. These embeddings, along with gender and age features, were used as input for the models LR, XGBoost, and EN.

In the training set, XGBoost acheived the best result with all embeddings, although this +result was not reflected in the test set, indicating potential overfitting. Despite variations in performance metrics, the EN model with BlueBERT embeddings performed best overall, with a specificity of 0.936, PPV of 0.464, AUPRC of 0.527, and F-1 of score 0.503. Overall, models using BERT-processed features, either BERT-base or BlueBERT, outperformed those using TF-IDF embeddings. The performance metrics of the models are summarized in Table 3.

Embedding	Model		AUROC	Sensitivity	Specificity	PPV	NPV	F-1	AUPRC	Accuracy
TF-IDF	LR	Train	0.714	0.5078	0.920	0.731	0.813	0.599	0.693	0.796
		Test	0.701	0.494	0.908	0.353	0.946	0.411	0.447	0.870
	XGB	Train	0.863	0.751	0.975	0.929	0.901	0.830	0.877	0.908
		Test	0.707	0.514	0.901	0.344	0.948	0.412	0.451	0.865
	EN	Train	0.639	0.319	0.959	0.769	0.767	0.451	0.646	0.767
		Test	0.646	0.337	0.955	0.434	0.934	0.379	0.416	0.898
BERT base	LR	Train	0.784	0.640	0.927	0.791	0.858	0.708	0.770	0.841
		Test	0.736	0.571	0.901	0.371	0.954	0.450	0.491	0.871
	XGB	Train	0.890	0.793	0.986	0.960	0.918	0.869	0.908	0.928
		Test	0.688	0.449	0.927	0.384	0.943	0.414	0.442	0.883
	EN	Train	0.753	0.575	0.932	0.783	0.837	0.663	0.743	0.825
		Test	0.730	0.539	0.921	0.408	0.952	0.465	0.495	0.885
BlueBERT	LR	Train	0.785	0.634	0.936	0.810	0.856	0.711	0.777	0.846
		Test	0.750	0.582	0.917	0.418	0.956	0.486	0.519	0.886
	XGB	Train	0.885	0.781	0.990	0.971	0.913	0.866	0.909	0.927
		Test	0.724	0.515	0.933	0.440	0.950	0.474	0.500	0.895

	EN	Train	0.757	0.571	0.944	0.813	0.837	0.671	0.756	0.832
		Test	0.742	0.549	0.936	0.464	0.953	0.503	0.527	0.900

Table 3. Performance metrics of trained models. Performance metrics of 3 models tested (LR, XGBoost, EN) on 3 different features generated by TF-IDF, BERT, BlueBERT respectively, are described in this table.

Feature Importance

To identify words significantly influencing the prediction of SAE presence in a patient, SHAP feature importance analysis was conducted. This analysis focused on the model with the best performance, which was EN model utilizing BlueBERT embeddings as features. Since this model included gender and age as features, a separate model was trained solely on text embeddings to isolate the impact of words only.

Following model training, SHAP analysis was performed. In SHAP analysis, each feature (token) is assigned with a specific SHAP value for each prediction. Cumulative SHAP values are the sum of SHAP values from all predictions in the dataset, resulting in a value indicating that specific feature's overall importance in making a prediction in the model.

To refine the results and focus on medically relevant information, non-medical vocabularies were excluded from the final list. Table 4 represents the result of this analysis, and Figure 4 is the visualization of the result for comparison.

No.	Token	Cumulative SHAP Value
1	pain	259.882050
2	injection	212.259079
3	site	196.070557
4	dose	174.045105
5	arm	170.956924
6	fever	115.826820
7	red	97.875015
8	rash	95.138184
9	headache	87.777504
10	left	71.082741
11	swelling	70.425987
12	chill	67.893684
13	sore	67.301796
14	symptoms	64.831779
15	body	51.241306

16	center	46.285011
17	muscle	44.791679
18	fatigue	43.881763
19	follow	42.231720
20	severe	41.073921

Table 4. Feature importance results. This table represents top 20 tokens with highest cumulative SHAP values from BlueBERT embeddings after filtering out medically irrelevant vocabularies.

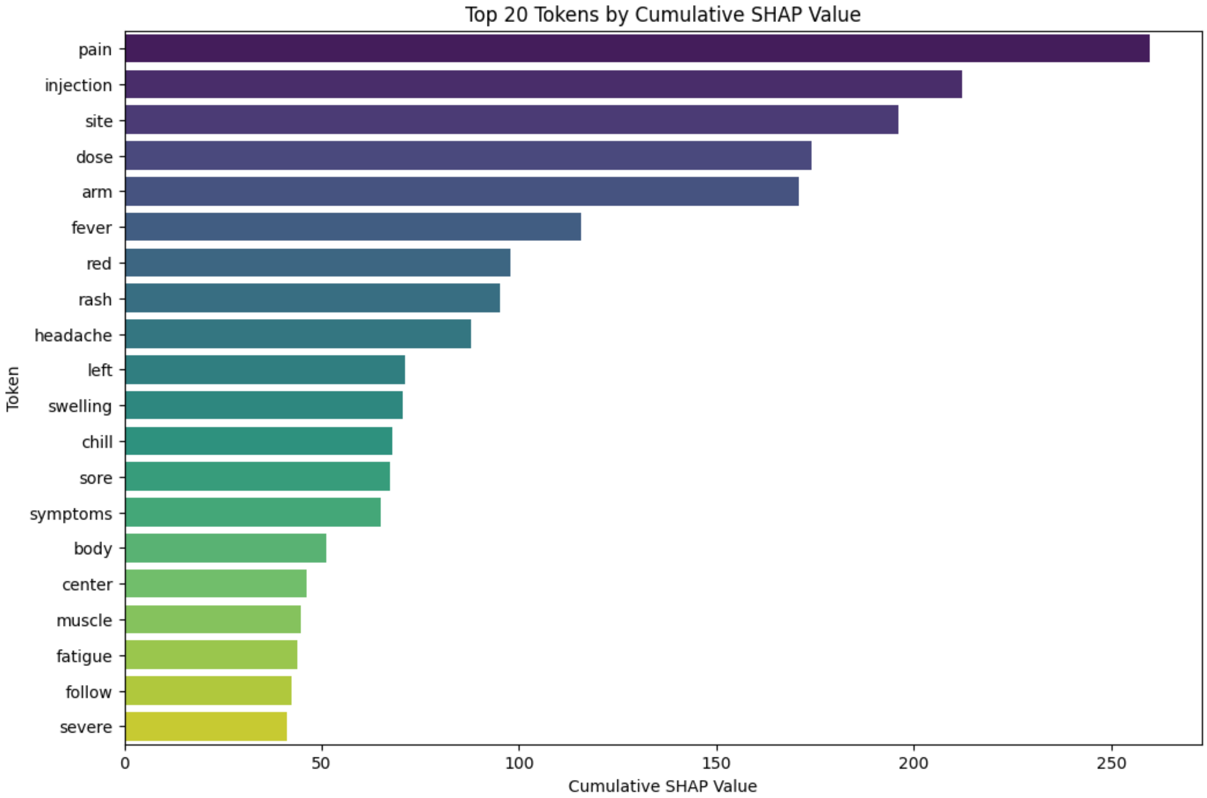

Fig. 4. Bar graph of feature importance result.

Subgroup Statistical Analysis

After cleaning data using the same method as in model training preparation and removing rows lacking both 'CUR_ILL' and 'HISTORY' values, a total of 27,282 rows were included for subgroup analysis. Number of counts for each condition groups as defined by ICD codes, as well as age groups are described in Table 1 and Table 2, respectively. Table 5 describes the number of counts from this dataset of disease entities identified by ScispaCy.

No.	Disease Entity	Frequency
1	hypertension	1,746
2	diabetes	1,065
3	asthma	1,021
4	hyperlipidemia	893
5	hypothyroidism	807
6	arthritis	678
7	depression	658
8	anxiety	507
9	osteoarthritis	471
10	shingles	439
11	pain	429
12	osteoporosis	382
13	allergies	362
14	fibromyalgia	316
15	hypothyroid	281
16	hypercholestrolemia	272
17	migraines	237
18	obesity	230
19	breast cancer	215
20	insomnia	212

Table 5. Top 20 frequency of disease entities These are 20 disease entities identified and extracted by ScispaCy with the highest frequency from the dataset. While the purpose of this extraction was to evaluate the relevance between pre-existing medical condition of patients and SAE, this list includes symptoms or result of AE, such as 'pain', 'migraines', and 'shingles'.

After calculating odds ratios using logistic regression, ICD code groups R00-R99, K00-K95, I00-I99, C00-D49, E00-E89, F01-F99, and G00-G99 showed the highest odds ratios with statistically significant p-values. This indicates that patients with conditions in these ICD code groups may have higher chance of experiencing SAE compared to those without any pre-existing medical conditions (Figure 5). For age group analysis, patients with age 75 or higher showed higher odds compared to age group 50-54 of experiencing SAE with statistically significant p-values (Figure 6). When we see a simple age group SAE ratio, we can also observe increasing SAE ratio with age (Figure 7). Between gender groups, female patients had significantly lower odds compared to that of the male group in this dataset (Figure 8). Tables 6, 7, and 8 illustrates overall results of odds ratio calculations for each of the categories.

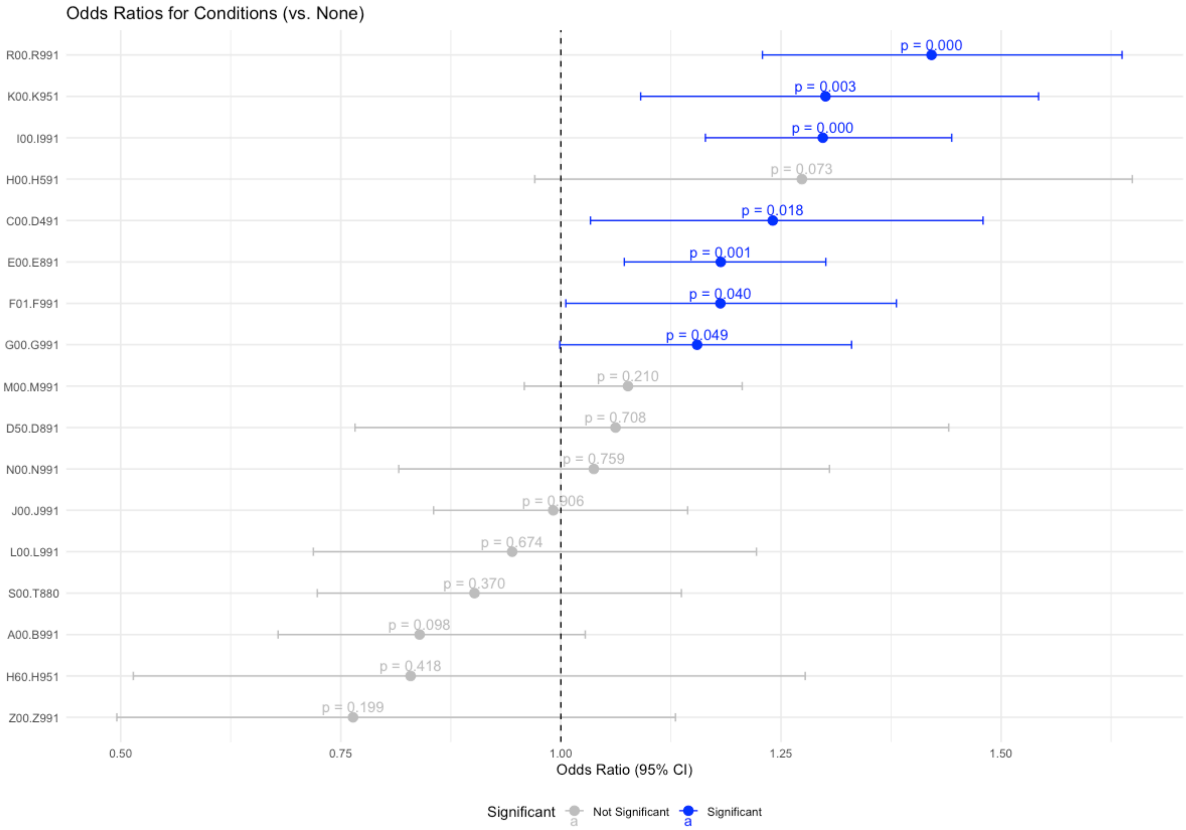

Fig. 5. Odds ratios for ICD code groups Odds ratio calculation along with its Confidence Interval (CI) values and p-values are described, with blue representing ICD code groups with statistically significant difference in odds compared to the patient group with no medical conditions.

ICD Group	Group Description	Odds Ratio	Std. Error	Statistic	P-Value	CI (Low)	CI (High)
A00 - B99	Certain infections and parasitic diseases	0.840	0.106	-1.655	0.098	0.679	1.028
C00 - D49	Neoplasms	1.241	0.091	2.358	0.018	1.034	1.480
D50 - D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanisms	1.062	0.161	0.375	0.708	0.766	1.441
E00 - E89	Endocrine, nutritional and metabolic diseases	1.182	0.049	3.381	0.001	1.072	1.301
F01 - F99	Mental, behavioral and neurodevelopmental disorders	1.181	0.081	2.059	0.040	1.006	1.381
G00 - G99	Diseases of the nervous system	1.155	0.073	1.968	0.049	0.999	1.330

H00 - H59	Diseases of the eye and adnexa	1.274	0.135	1.791	0.073	0.970	1.649
H60 - H95	Diseases of the ear and mastoid process	0.829	0.231	-0.810	0.418	0.514	1.278
I00 - I99	Diseases of the circulatory system	1.298	0.055	4.742	0.000	1.164	1.444
J00 - J99	Diseases of the respiratory system	0.991	0.074	-0.118	0.906	0.855	1.144
K00 - K95	Diseases of the digestive system	1.301	0.088	2.973	0.003	1.091	1.543
L00 - L99	Diseases of the skin and subcutaneous tissue	0.945	0.135	-0.421	0.674	0.719	1.222
M00 - M99	Diseases of the muscoskeletal system and connective tissue	1.076	0.059	1.254	0.210	0.958	1.206
N00 - N99	Diseases of the genitourinary system	1.037	0.120	0.307	0.759	0.816	1.305
R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	1.421	0.073	4.801	0.000	1.229	1.637
S00-T88	Injury, poisoning and certain other consequences of external causes	0.902	0.115	-0.896	0.370	0.723	1.137
Z00-Z99	Factors influencing health status and contact with health services	0.764	0.210	-1.285	0.199	0.496	1.130

Table 6. ICD Code Group Analysis Results This table presents the odds ratios for each ICD code groups compared to the 'none' group, which represents patients with no medical conditions or history.

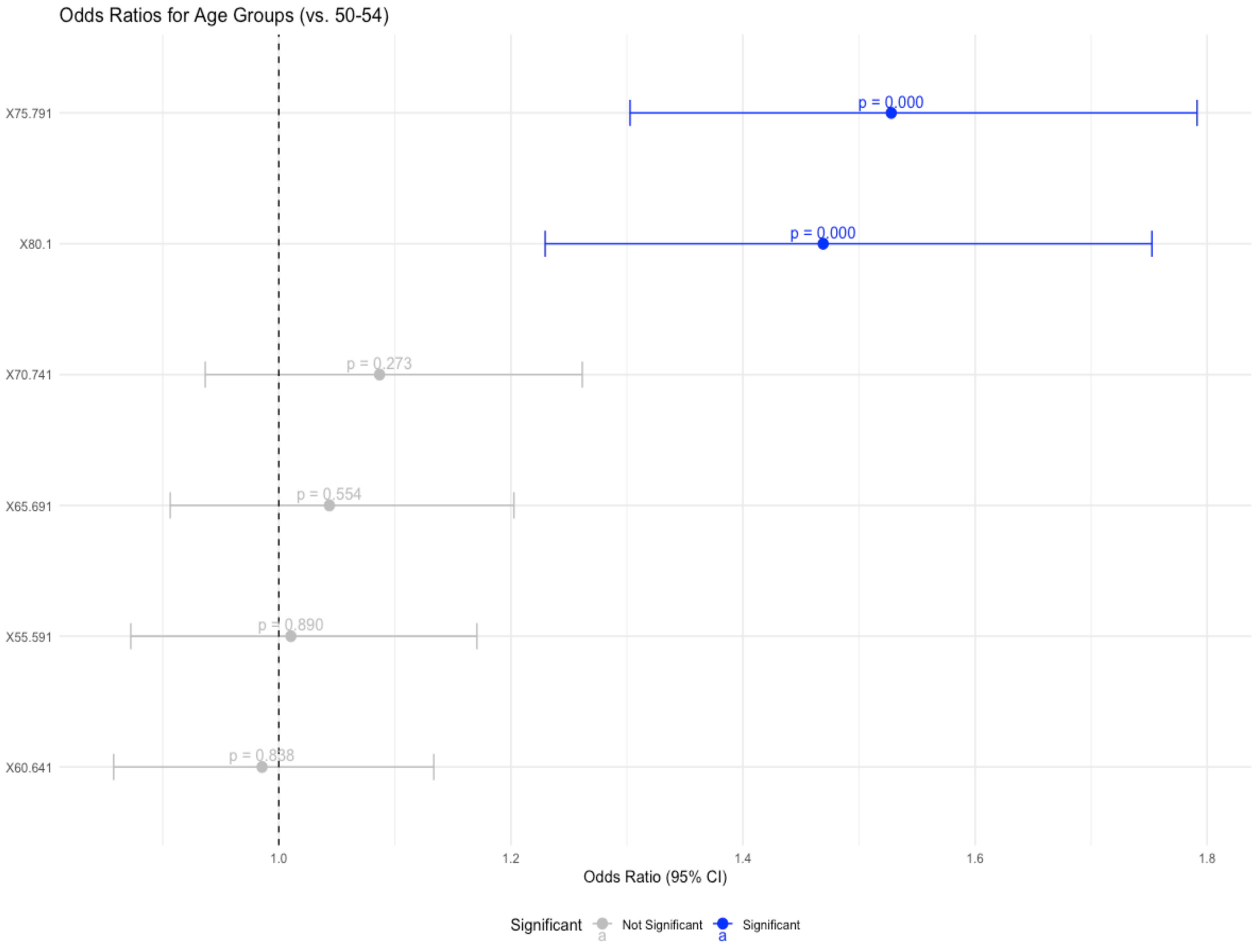

Fig. 6. Odds ratios for age groups Age groups with statistically different odds compared to reference (age 50-54) group are represented in blue colors.

Age Group	Odds Ratio	Std. Error	Statistic	P-Value	CI (Low)	CI (High)
55-59	1.010	0.075	0.138	0.890	0.872	1.171
60-64	0.986	0.071	-0.205	0.838	0.858	1.134
65-69	1.044	0.072	0.592	0.554	0.906	1.203
70-74	1.087	0.076	1.096	0.273	0.937	1.262
75-79	1.528	0.081	5.217	0.000	1.303	1.791
80+	1.469	0.090	4.257	0.000	1.230	1.752

Table 7. Age Group Analysis Results This table presents the odds ratios for each age groups compared to the '50-54' age group.

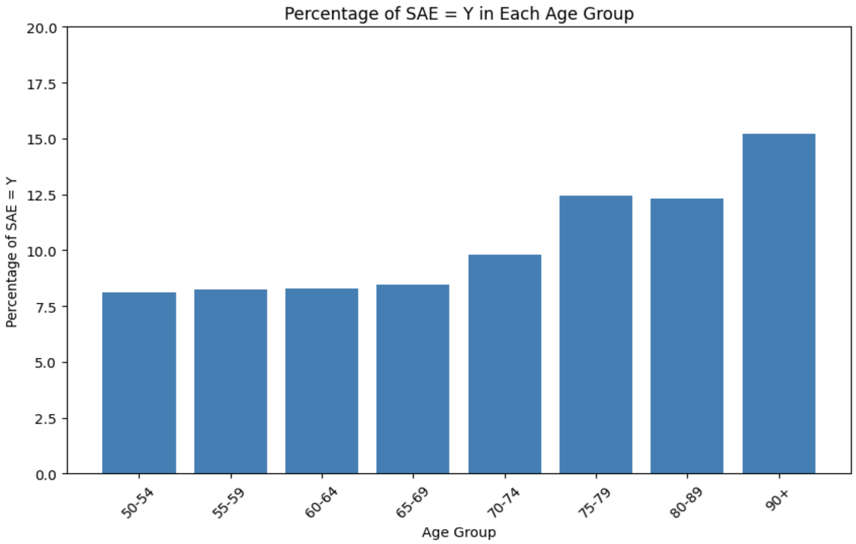

Fig. 7. Percentage of SAE in each age group Ratio of patients who resulted in SAE from each age group is described in this bar graph. For the purpose of subgroup analysis in this project, patients at the age of 80 or older were combined into one age group.

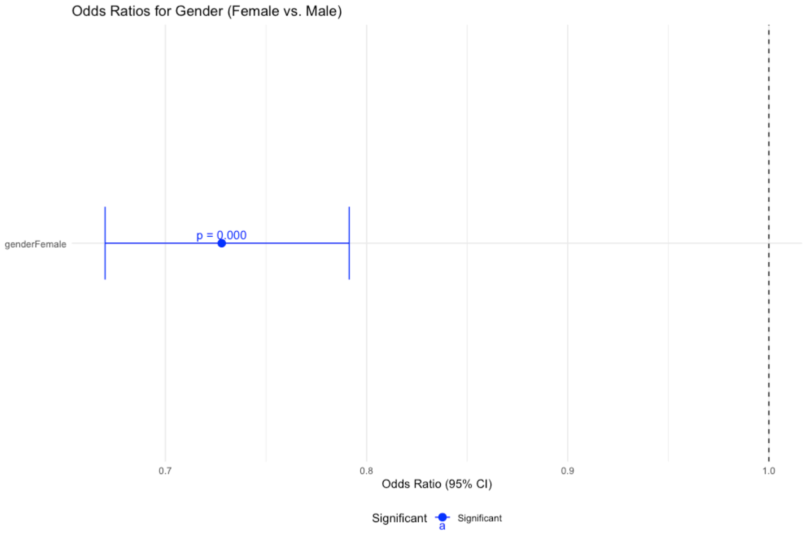

Fig. 8. Odds ratios for gender In this figure, odds of female SAE occurrences were compared to those of males.

Gender	Odds Ratio	Std. Error	Statistic	P-Value	CI (Low)	CI (High)
Female	0.727	0.042	-7.515	0.000	0.669	0.790

Table 8. Gender Group Analysis Results This table presents the odds ratios for females compared to male patients.

Discussion

Confidence in vaccines is crucial for overcoming vaccine hesitancy, particularly among elderly population. Reliable data is essential for making informed healthcare decision. In this study, VAERS dataset was utilized to develop a predictive model for SAEs, focusing on symptoms, medical conditions, and demographics. TF-IDF, BERT and BlueBERT models were used for text embeddings, and LR, XGBoost, as well as EN models were trained. EN model with BlueBERT embeddings demonstrated the most promising performance overall. BlueBERT is a model pre-trained on PubMed abstracts and clinical notes from MIMIC-III, which may be the factor why this model was able to capture medical contexts better compared to BERT-base model. While the XGBoost model showed potential, overfitting was observed, indicating the challenges of high-dimensional data and imbalanced dataset.

Due to class-imbalance problem with the data, we aimed to improve AUPRC as this metric assesses the trade-off between precision and recall, which aligns with the objective of identifying SAE occurrence possibility despite its scarcity. The highest AUPRC achieved was 0.527, which indicates potential for improvement.

Following model development and performance evaluation, feature importance was conducted on model with the best overall performance metrics, which was the EN model with BlueBERT embeddings. SHAP analysis highlighted the importance of specific medical terms and symptoms in SAE prediction such as pain, fever, red, rash, and headache, aligning with existing knowledge of possible AE symptoms according to Shingrix clinical data. [20]

For subgroup analysis, the 'CUR_ILL' and 'HISTORY' columns were processed, and ScispaCy was used to extract 'DISEASE' entities from the text, in order to group conditions for analysis. After identifying disease entities and mapping these with ICD-10 codes, these were grouped based on similar symptoms for analysis. As for age, age groups were defined after the assessment distribution of age in the dataset. ICD code groups, age groups, along with gender were included in the logistic regression analysis to calculate odds ratios and assess their combined effects on different categories. R00-R99 (Symptoms, signs and abnormal clinical and laboratory findings), I00-I99 (Diseases of the circulatory system), K00-K95 (Diseases of the digestive system), C00-D49 (Neoplasms), E00-E89 (Endocrine, nutritional and metabolic diseases), F01-F99 ('Mental, behavioral and neurodevelopmental disorder), and G00-G99 (Diseases of the nervous system) showed higher odds ratios with significant p-values. R00-R99 group primarily comprises symptom categories such as rash (R21), which may be indicative of adverse event description rather than pre-existing conditions. In comparison, other diseases areas include diseases such as hypertension, diabetes, and cancer which have high number counts in the dataset. In terms of the age group analysis, older age population (75 years and older) showed statistically significantly higher odds compared to 50-54 age group in terms of SAE prevalence. This finding also aligns with the result from the conditions group with the increased risk of certain conditions in older age groups such as hypertension, diabetes, and depression.[21-23] As for gender, female had significantly lower odds compared to male of SAE occurrence.

Limitations

While this study demonstrates the potential of utilizing the VAERS dataset to develop prediction models and provide information beneficial for those vaccine recipients, it does have limitations due to the nature of the dataset. VAERS is comprised of reports submitted by the general public, and not limited to healthcare professionals, and therefore, the content of the report is not regulated which affects the quality and consistency of the dataset. For instance, the 'CUR_ILL' and 'HISTORY' columns both contain current and past medical conditions of the patients. In a lot of cases, symptoms due to AE is also described in 'CUR_ILL' section, which overlap may explain the highest odds ratio in the R00-R99 category which highest counts condition includes 'pain', and similarly with G00-G99, which includes 'migraine'. Consequently, it is difficult to distinguish pre-existing conditions from AE symptoms within the dataset.

Additionally, since the reports were submitted after the occurrence of SAEs, there might be information within the 'SYMPTOM_TEXT' directly indicating SAE, potentially influencing the model's prediction capabilities. While efforts were made to minimize this factor by eliminating certain words directly indicating SAE for TF-IDF embeddings, the underlying issue remains. Despite these limitations, this study highlights the potential value for further research in this area. By refining data collection, processing, and analysis methods. Future studies could overcome these challenges and provide even more reliable insights to inform healthcare providers and patients.

Conclusion

This study has the potential to predict the progression of SAEs at the onset of AEs by analyzing patient's symptom description, medical history and condition, and demographics information. The ability to predict SAE progression at the time of AE onset enables proactive patient care and intervention before SAEs occur. Furthermore, our statistical analysis identified specific medical conditions, age groups, and gender as factors associated with a higher probability of developing SAEs following AE occurrences. With this information, healthcare providers and patients can have a more informed and personalized decision when it comes to how to manage after onset of AE during their early stages.

References

1. (CDC), C.f.D.C.a.P. *Shingles (Herpes Zoster) Vaccination*. n.d. June 24, 2024]; Available from: https://www.cdc.gov/vaccinesafety/vaccines/shingles-herpes-vaccine.html.
2. (CDC), C.f.D.C.a.P. *Shingrix*. n.d. June 24, 2024]; Available from: https://www.cdc.gov/vaccines/vpd/shingles/public/shingrix/index.html.
3. Wang, Q., et al., *Willingness to Vaccinate Against Herpes Zoster and Its Associated Factors Across WHO Regions: Global Systematic Review and Meta-Analysis.* JMIR Public Health Surveill, 2023. **9**: p. e43893.
4. Gavi, t.V.A. *Vaccine hesitancy is one of the greatest threats to global health – and the pandemic has made it worse*. n.d. July 24, 2024]; Available from: https://www.gavi.org/vaccineswork/vaccine-hesitancy-one-greatest-threats-global-health-and-pandemic-has-made-it-worse.
5. (FDA), F.a.D.A. *Reporting Serious Problems to FDA*. n.d. June 24, 2024]; Available from: https://www.fda.gov/safety/reporting-serious-problems-fda/what-serious-adverse-event.
6. (WHO), W.H.O., *The conceptual framework for the international classification for patient safety.* 2009.
7. Turing. *Guide on Word Embeddings in NLP*. July 24, 2024]; Available from: https://www.turing.com/kb/guide-on-word-embeddings-in-nlp.
8. Ramos, J.E. *Using TF-IDF to Determine Word Relevance in Document Queries*. 2003.
9. Devlin, J., Chang, M., Lee, K., & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. in *North American Chapter of the Association for Computational Linguistics (NAACL)*. 2019.
10. Peng, Y., S. Yan, and Z. Lu, *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*, in *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019*. 2019, Association for Computational Linguistics.
11. Tabaie, A., et al., *Integrating structured and unstructured data for timely prediction of bloodstream infection among children.* Pediatr Res, 2023. **93**(4): p. 969-975.
12. scikit-learn. *StandardScaler*. July 24, 2024]; Available from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.
13. Sperandei, S., *Understanding logistic regression analysis.* Biochemia Medica, 2014. **24**(1): p. 12-18.
14. Moore, A.B., M., *XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study.* Clinical Medicine Insights: Cardiology, 2022. **16**.
15. Zou, H., Hastie, T., *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. **67**(2): p. 301-320.
16. Lundberg, S.M., & Lee, S. I. . *SHAP (SHapley Additive exPlanations)*. 2017; Available from: https://github.com/shap/shap.
17. Murty, S., et al., *Hierarchical Losses and New Resources for Fine-Grained Entity Typing and Linking*, in *Proceedings of the 2018 Annual Meeting of the Association for Computational Linguistics (ACL 2018)*. 2018, Association for Computational Linguistics (ACL). p. 104-114.
18. Neumann, M., et al., *SciSpacy: Fast and Robust Models for Biomedical Natural Language Processing*, in *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*. 2019, Association for Computational Linguistics (ACL). p. 45-53.
19. ICD10Data.com, *ICD-10-CM Codes*.
20. (FDA), F.a.D.A. *Shingrix Prescribing Information*. 2023 [cited July 25, 2024; Available from: https://www.fda.gov/media/108597/download.
21. Leszczak, J., et al., *Risk factors and prevalence of hypertension in older adults from south-eastern Poland: an observational study.* Scientific Reports, 2024. **14**.

22. Yan, Z.C., M.; Han, X.; Chen, Q.; Lu, H., *The Interaction Between Age and Risk Factors for Diabetes and Prediabetes: A Community-Based Cross-Sectional Study.* International Journal of Environmental Research and Public Health, 2023. **20**(6).

23. Nakua, E.K.A., J.; Tawiah, P.; Barnie, B.; Donkor, P.; Mock, C., *The prevalence and correlates of depression among older adults in Greater Kumasi of the Ashanti Region.* BMC Geriatrics, 2022. **22**.