GEORGETOWN UNIVERSITY

MedStar Health

Predictive Modeling and Analysis of Adverse Events Post-Shingles Vaccination

: Serious Adverse Event Prediction Model Development & Data Analysis Using VAERS Dataset

Youngmin Choi (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)

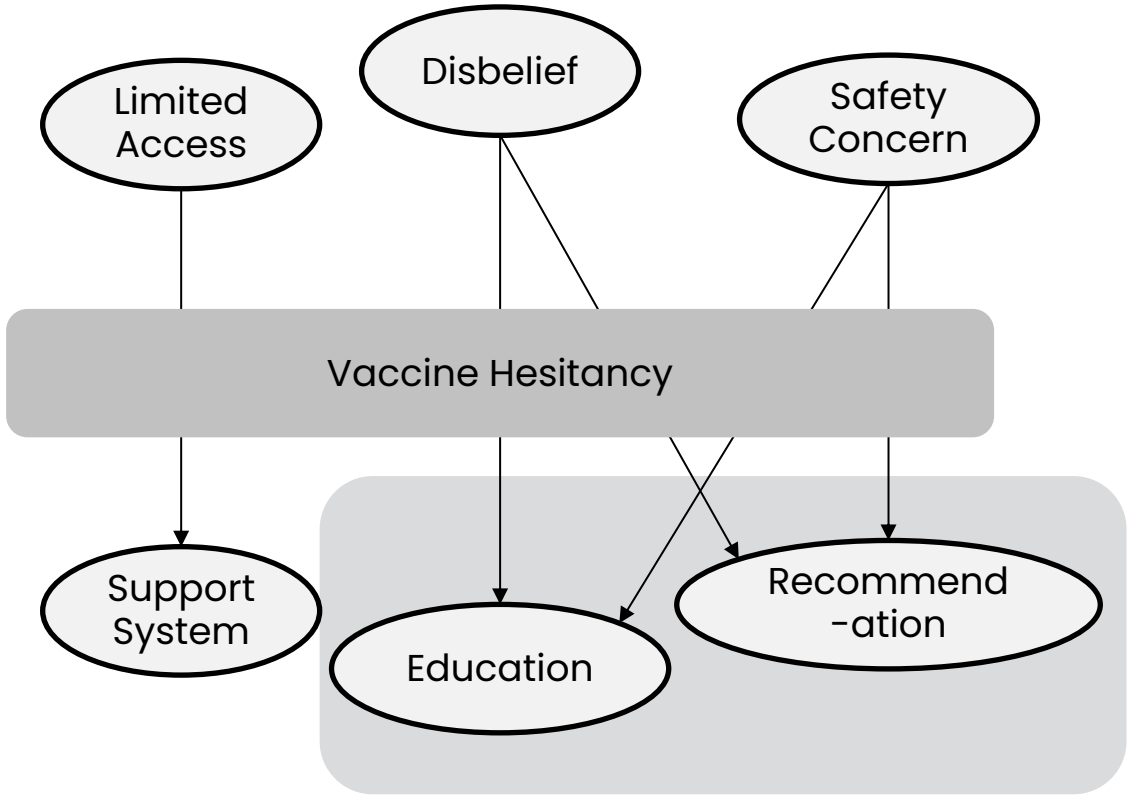
The Washington Post
Democracy Dies in Darkness

Opinion | Vaccine hesitancy is growing. Doctors will have to fight back.

A nurse prepares a coronavirus vaccine dose at the Ward 4 D.C. Covid Center on March 31. (Eric Lee for The Washington Post)

By the Editorial Board

May 17, 2023 at 4:53 p.m. EDT



PROJECT INTRODUCTION

Problem	• Concerns of potential adverse events may lead to hesitancy among potential recipients. • Shingles vaccines are not mandatory, so individual decision-making is important.
Scope	• Develop SAE (after AE onset) prediction model using VAERS data. • Conduct statistical analysis of subgroups.
Goal	• Target AUPRC improvement of 20% (vs baseline model). • Identify groups with higher risks.
Benefits	• Early intervention can lead to better patient care outcomes and potentially reduce healthcare costs. • Patients and healthcare professionals can make more confident and informed decisions.

Timeline

May 2024 – August 2024

Method

- Python
- R
- NLP
- Machine Learning
- Statistical Analysis

Sponsor

MedStar Health

Industry Mentor

Dr. Azade Tabie, Data Scientist, MedStar Health

Academic Mentor

Dr. Peter McGarvey, Director, ICBI, Georgetown University

TABLE OF CONTENTS

01 INTRODUCTION

SHINGLES & VACCINATION

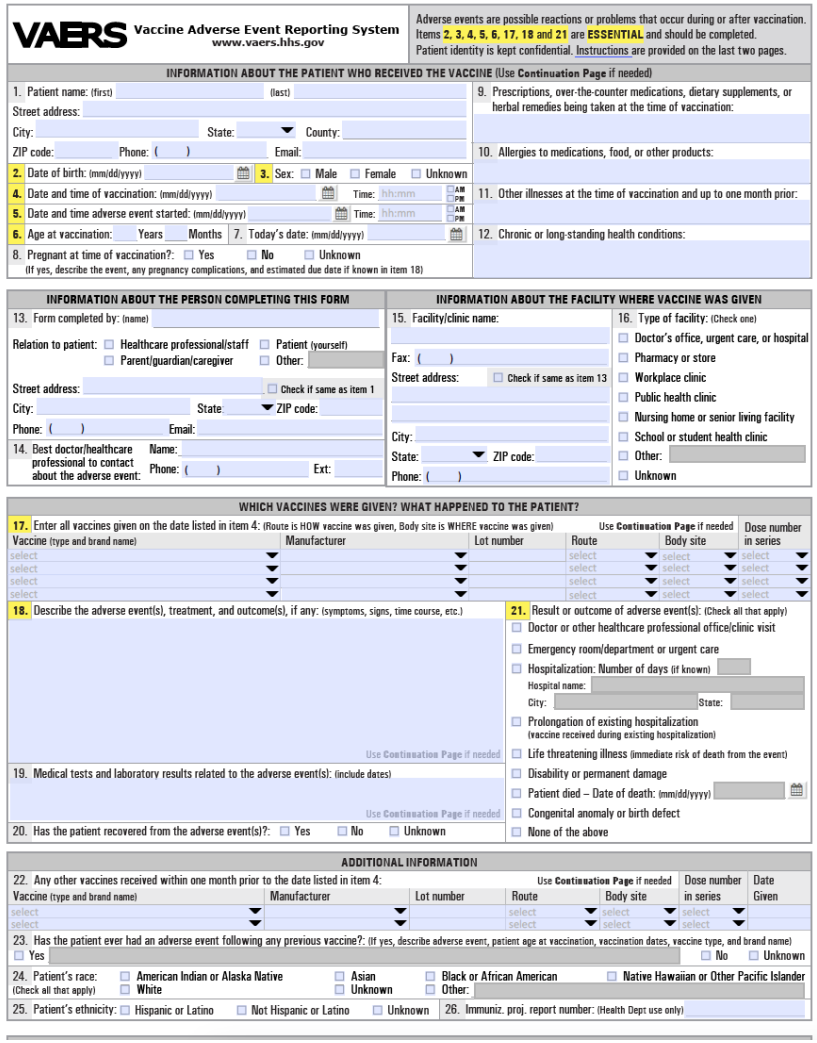
What is Shingles?

- Caused by reactivation of dormant Varicella-Zoster Virus.
- 1/3 people in US will have shingles in their lifetime.
- Postherpetic neuralgia (PHN).

Shingles Vaccine

- Recombinant Shingles Vaccine approved by FDA in 2017 (Shingrix).
- Adults aged 50 years and older (2 injections)

VAERS DATA & SAE



VAERS Vaccine Adverse Event Reporting System
www.vaers.hhs.gov

Adverse events are possible reactions or problems that occur during or after vaccination. Items 2, 3, 4, 5, 6, 17, 18 and 21 are ESSENTIAL and should be completed. Patient identity is kept confidential. Instructions are provided on the last two pages.

INFORMATION ABOUT THE PATIENT WHO RECEIVED THE VACCINE (Use Continuation Page if needed)

1. Patient name: (first) (last)
Street address:
City: State: County:
ZIP code: Phone: ( ) Email:
2. Date of birth: (mm/dd/yyyy) 3. Sex: Male Female Unknown
4. Date and time of vaccination: (mm/dd/yyyy) Time: hh:mm AM PM
5. Date and time adverse event started: (mm/dd/yyyy) Time: hh:mm AM PM
6. Age at vaccination: Years Months 7. Today's date: (mm/dd/yyyy)
8. Pregnant at time of vaccination?: Yes No Unknown
(If yes, describe the event, any pregnancy complications, and estimated due date if known in item 18)
9. Prescriptions, over-the-counter medications, dietary supplements, or herbal remedies being taken at the time of vaccination:
10. Allergies to medications, food, or other products:
11. Other illnesses at the time of vaccination and up to one month prior:
12. Chronic or long-standing health conditions:

INFORMATION ABOUT THE PERSON COMPLETING THIS FORM

13. Form completed by: (name)
Relation to patient: Healthcare professional/staff Patient (yourself) Parent/guardian/caregiver Other:
Street address: Check if same as item 1
City: State: ZIP code:
Phone: ( ) Email:
14. Best doctor/healthcare professional to contact about the adverse event: Name: Phone: ( ) Ext:

INFORMATION ABOUT THE FACILITY WHERE VACCINE WAS GIVEN

15. Facility/clinic name:
Fax: ( )
Street address: Check if same as item 13
City: State: ZIP code:
Phone: ( )
16. Type of facility: (Check one)
Doctor's office, urgent care, or hospital
Pharmacy or store
Workplace clinic
Public health clinic
Nursing home or senior living facility
School or student health clinic
Other:
Unknown

WHICH VACCINES WERE GIVEN? WHAT HAPPENED TO THE PATIENT?

17. Enter all vaccines given on the date listed in item 4: (Route is HOW vaccine was given, Body site is WHERE vaccine was given) Use Continuation Page if needed

Vaccine (type and brand name)	Manufacturer	Lot number	Route	Body site	Dose number in series
select			select	select	select
select			select	select	select
select			select	select	select
select			select	select	select

18. Describe the adverse event(s), treatment, and outcome(s), if any: (symptoms, signs, time course, etc.)
Use Continuation Page if needed
19. Medical tests and laboratory results related to the adverse event(s): (include dates)
Use Continuation Page if needed
20. Has the patient recovered from the adverse event(s)?: Yes No Unknown
21. Result or outcome of adverse event(s): (Check all that apply)
Doctor or other healthcare professional office/clinic visit
Emergency room/department or urgent care
Hospitalization: Number of days (if known)
Hospital name:
City: State:
Prolongation of existing hospitalization (vaccine received during existing hospitalization)
Life threatening illness (immediate risk of death from the event)
Disability or permanent damage
Patient died – Date of death: (mm/dd/yyyy)
Congenital anomaly or birth defect
None of the above

ADDITIONAL INFORMATION

22. Any other vaccines received within one month prior to the date listed in item 4: Use Continuation Page if needed

Vaccine (type and brand name)	Manufacturer	Lot number	Route	Body site	Dose number in series	Date Given
select			select	select	select	
select			select	select	select	

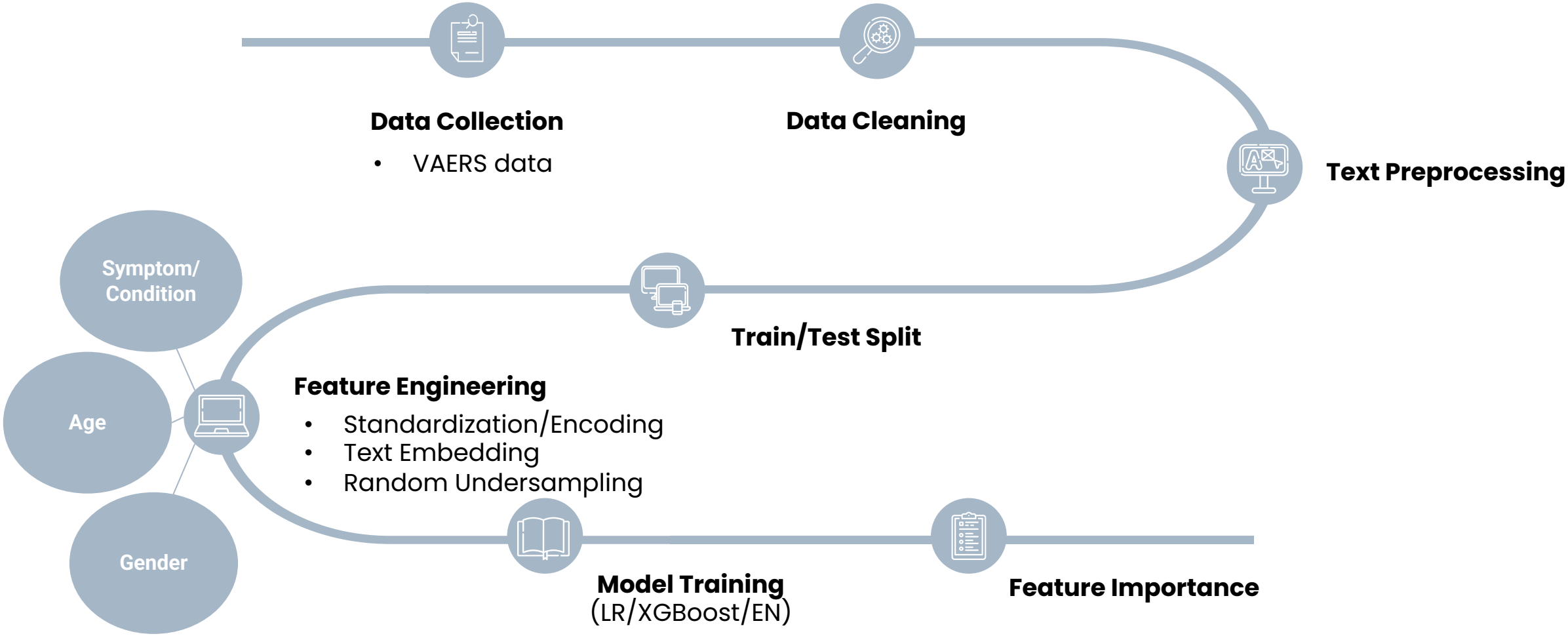
23. Has the patient ever had an adverse event following any previous vaccine?: (If yes, describe adverse event, patient age at vaccination, vaccination dates, vaccine type, and brand name)
Yes No Unknown
24. Patient's race: (Check all that apply) American Indian or Alaska Native Asian Black or African American Native Hawaiian or Other Pacific Islander White Unknown Other:
25. Patient's ethnicity: Hispanic or Latino Not Hispanic or Latino Unknown
26. Immuniz. proj. report number: (Health Dept use only)

- Vaccine Adverse Event Reporting System (VAERS) is a database of adverse event reports submitted by healthcare professionals, patients, caregivers, etc.
- No. reports submitted in 2023: 106,125 (48,438 in 2019)
- Data comprises information about vaccine administered, adverse events (symptom, outcome, etc.), and patient information.
- Serious Adverse Event (SAE) is an event with serious patient outcomes including death, ER visit, hospitalization and/or prolonged hospitalization, disability or permanent damage.

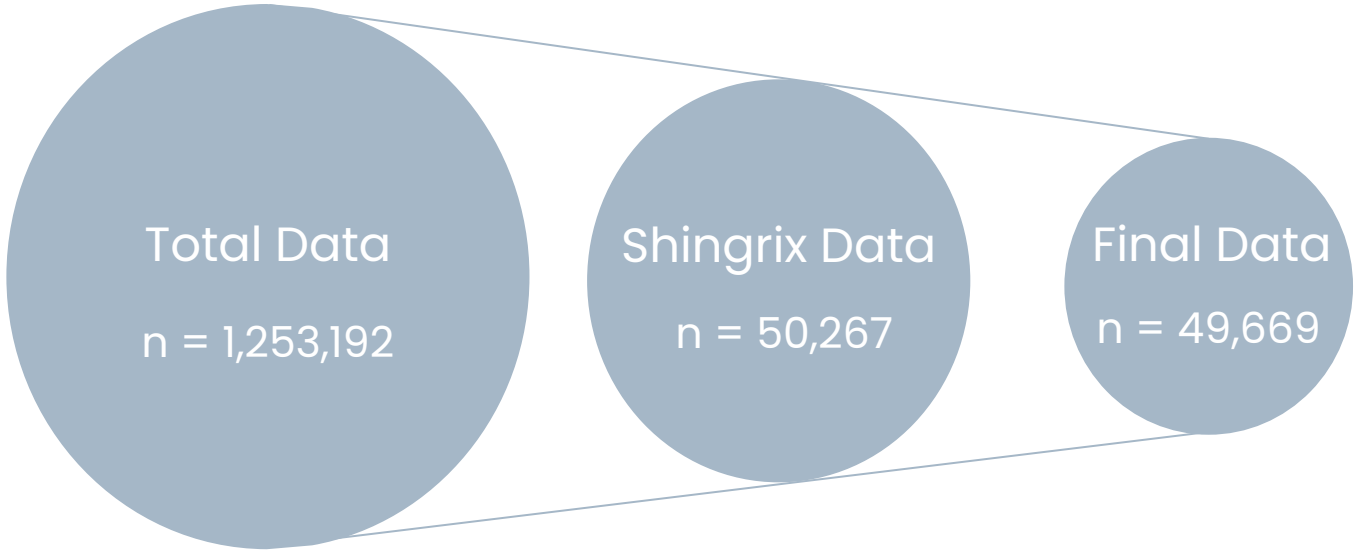
02 PREDICTIVE MODELLING - METHOD

Question:
When a patient starts to show signs of adverse events following vaccination, can we predict whether this will progress into serious adverse events?

PROCESS OVERVIEW



DATA EXPLORATION & CLEANING



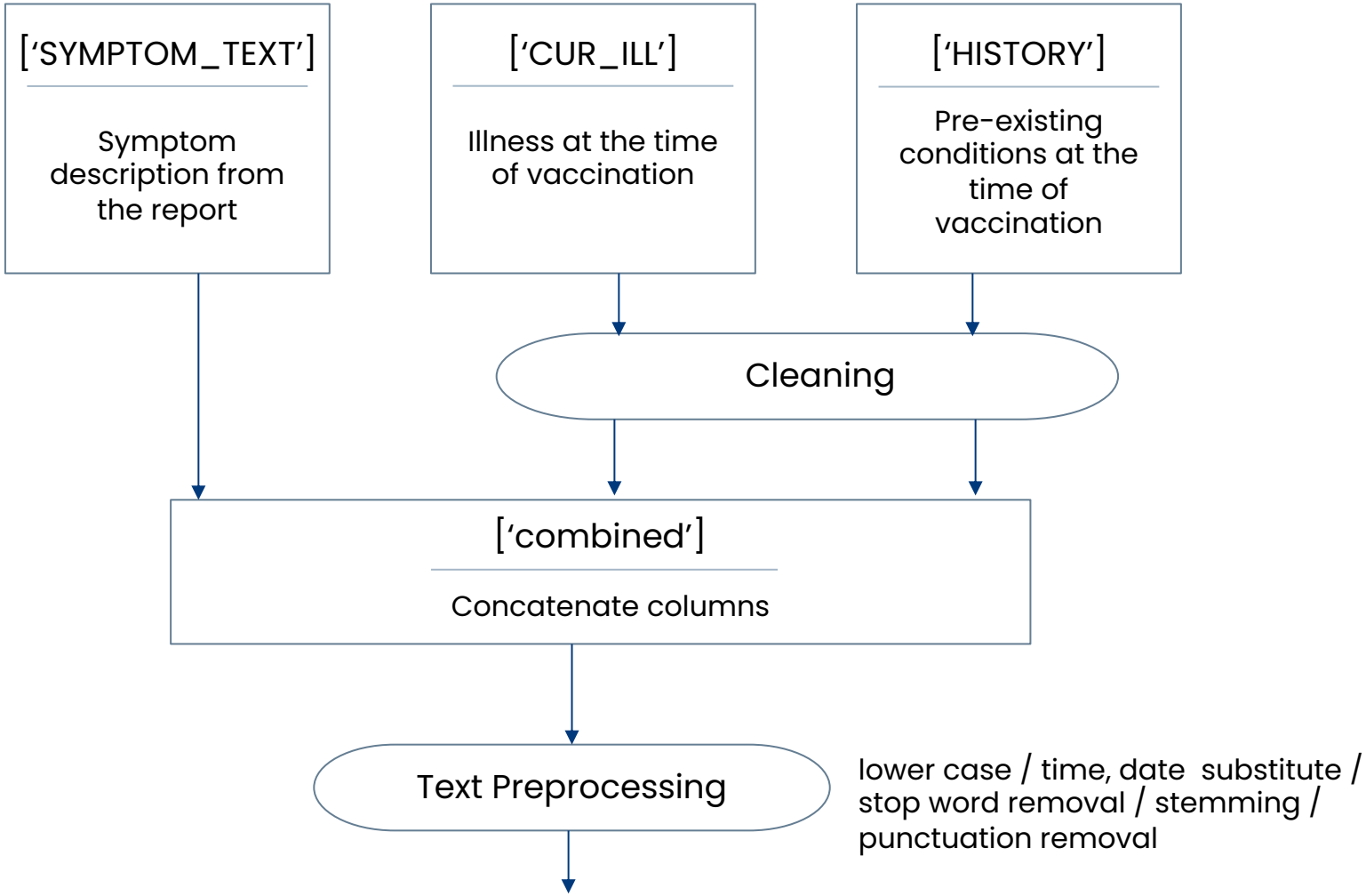
- From year 2018 to 2023
- Significant increase following covid vaccine
- 2018-2020 approx.
- 49,000 reports/year

- Vaccine : Shingrix
- Age : 50 years +

- Symptom NA removed
- Gender 'U' removed
- Symptom 'none stated' removed

- ✓ This study focused on data collected between **2018 and 2023** regarding Shingrix vaccine, which was approved by the FDA in 2017.
- ✓ The data included reports from patients aged **50 and older** who received the vaccine.
- ✓ To ensure data quality, rows with no symptom information and cases with unknown gender were removed from the analysis.
- ✓ From total final data of **49,669 reports**, 45,084 were classified as SAE and 4,585 as non-SAE, representing approximately **10%**.

TEXT PREPROCESSING



FEATURE ENGINEERING

Train/Test Split

- 80:20 ratio
- Stratified split

Text (combined)

TF-IDF BERT BlueBERT

- TFIDF (Term Frequency-Inverse Document Frequency)
- BERT
  (Bidirectional Encoder Representations from Transformers)
- BlueBERT

Age	Gender
StandardScaler	One-hot encoding

Random Undersampling

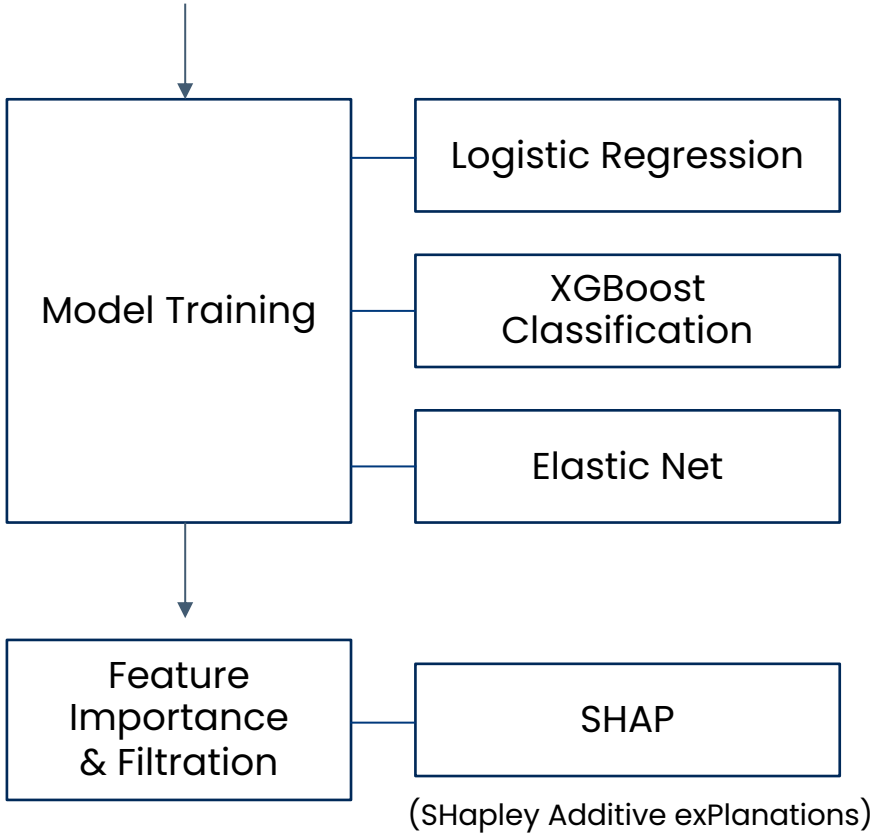
- Random sampling from non-SAE group for 7:3 ratio.

Train (n=40,175)
36,492 / 3,683

↓

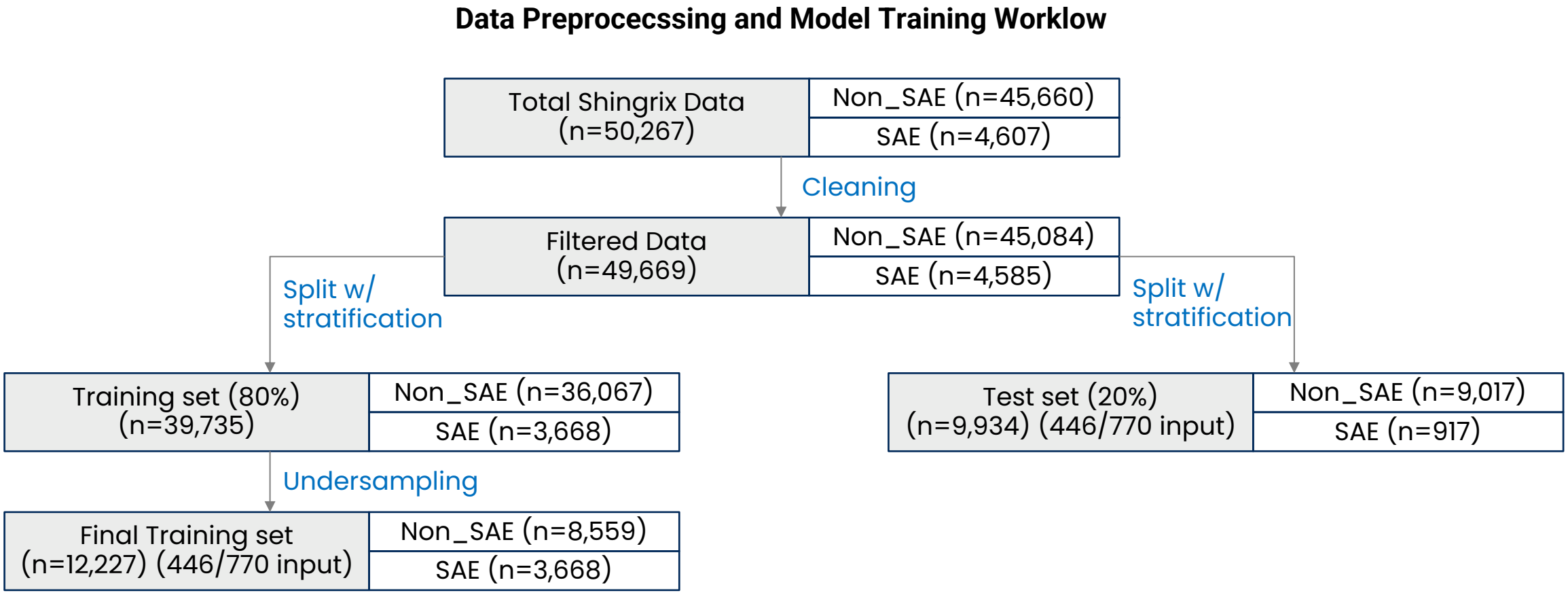
Train (n=12,277)
8,594 / 3,683

MODEL TRAINING & FEATURE IMPORTANCE

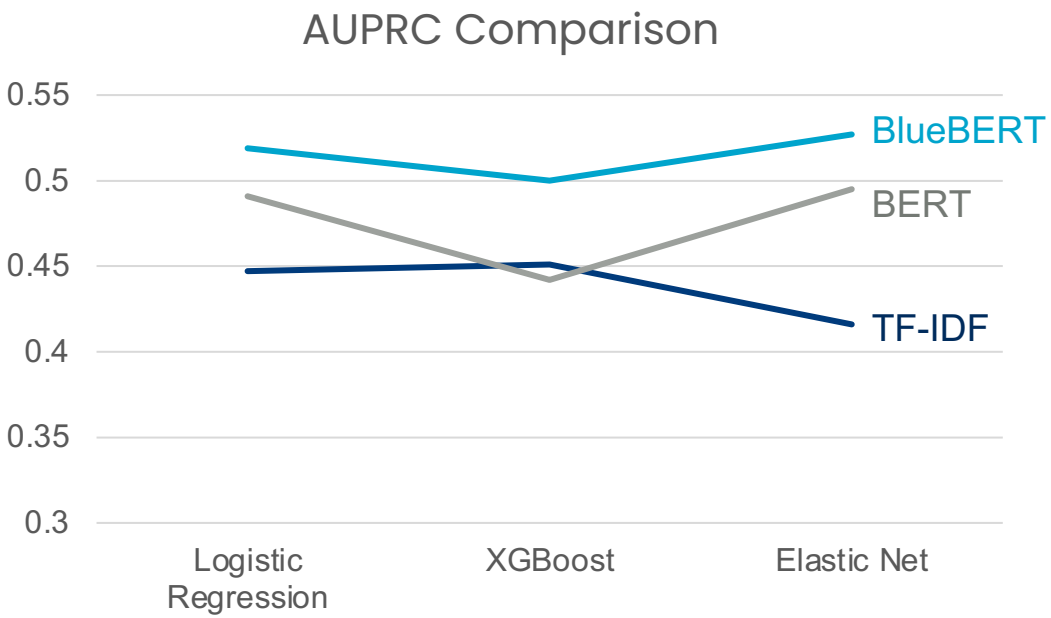


03 PREDICTIVE MODELLING - RESULTS

DATA PROCESSING RESULT



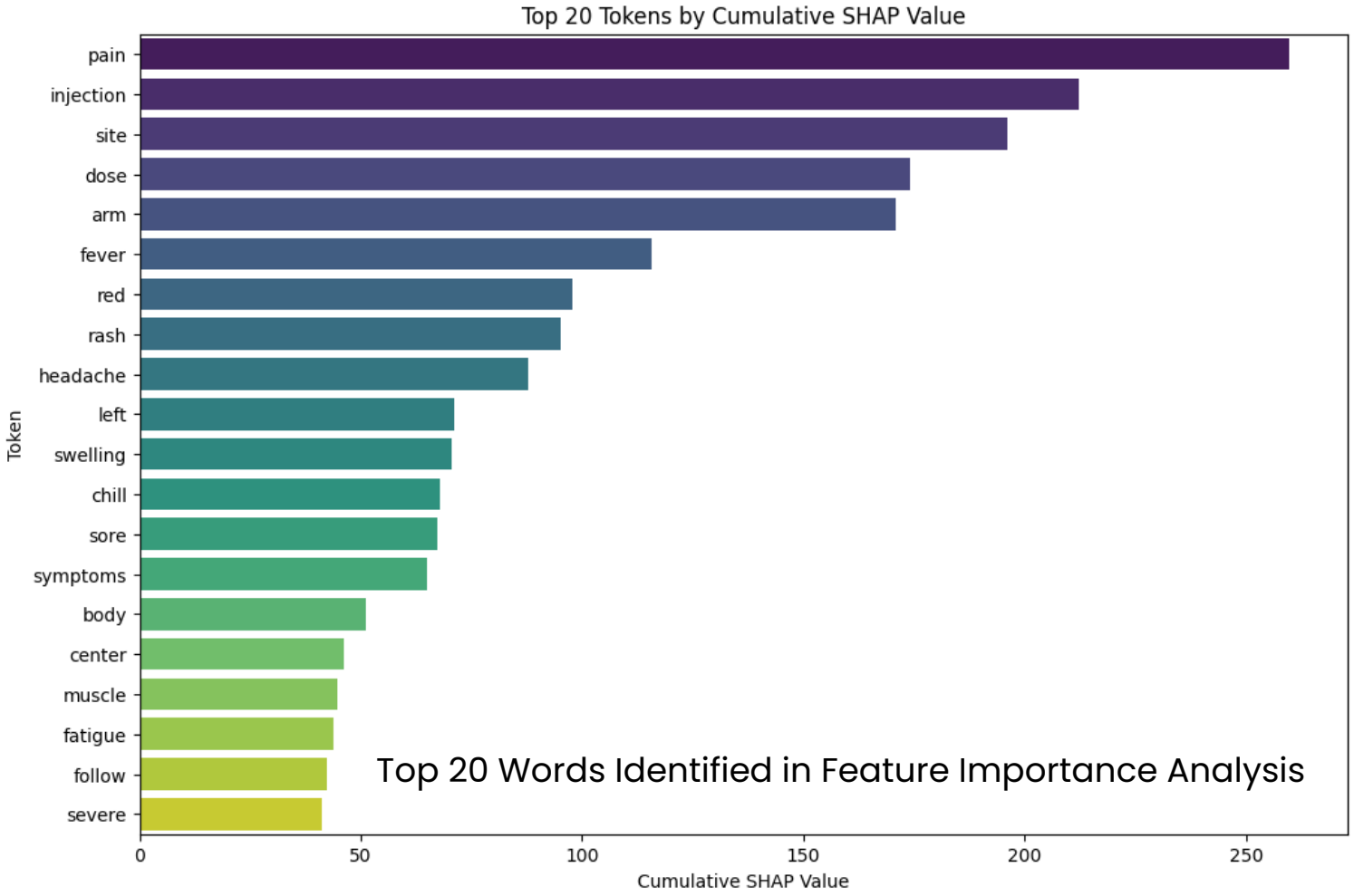
MODEL PERFORMANCE



Feature extraction	TF-IDF						BERT-base						BlueBERT					
Model	LR		XGBoost		EN		LR		XGBoost		EN		LR		XGBoost		EN	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
AUROC	0.714	0.701	0.863	0.707	0.639	0.646	0.784	0.736	0.890	0.688	0.753	0.73	0.785	0.750	0.885	0.724	0.757	0.742
Sensitivity	0.507	0.494	0.751	0.514	0.319	0.337	0.64	0.571	0.793	0.449	0.575	0.539	0.634	0.582	0.781	0.515	0.571	0.549
Specificity	0.92	0.908	0.975	0.901	0.959	0.955	0.927	0.901	0.986	0.927	0.932	0.921	0.936	0.917	0.990	0.933	0.944	0.936
PPV	0.731	0.353	0.929	0.344	0.769	0.434	0.791	0.371	0.960	0.384	0.783	0.408	0.810	0.418	0.971	0.440	0.813	0.464
NPV	0.813	0.946	0.901	0.948	0.767	0.934	0.858	0.954	0.918	0.943	0.837	0.952	0.856	0.956	0.913	0.950	0.837	0.953
F-1 score	0.599	0.411	0.83	0.412	0.451	0.379	0.708	0.45	0.869	0.414	0.663	0.465	0.711	0.486	0.866	0.474	0.671	0.503
AUPRC	0.693	0.447	0.877	0.451	0.646	0.416	0.77	0.491	0.908	0.442	0.743	0.495	0.777	0.519	0.909	0.500	0.756	0.527
Accuracy	0.796	0.87	0.908	0.865	0.767	0.898	0.841	0.871	0.928	0.883	0.825	0.885	0.846	0.886	0.927	0.895	0.832	0.900

- Across various text feature extraction and model combinations, BlueBERT outperformed TF-IDF and BERT-base, demonstrating superior capability in the biomedical domain.
- BlueBERT processed feature with elastic net model (C=0.1, ratio=0.3) acheived the highest metrics, indicating enhanced performance in identifying positive cases.
- Compared to baseline (TF-IDF/LR model) the performance improved by 18%.

FEATURE IMPORTANCE



Top 20 Words Identified in Feature Importance Analysis

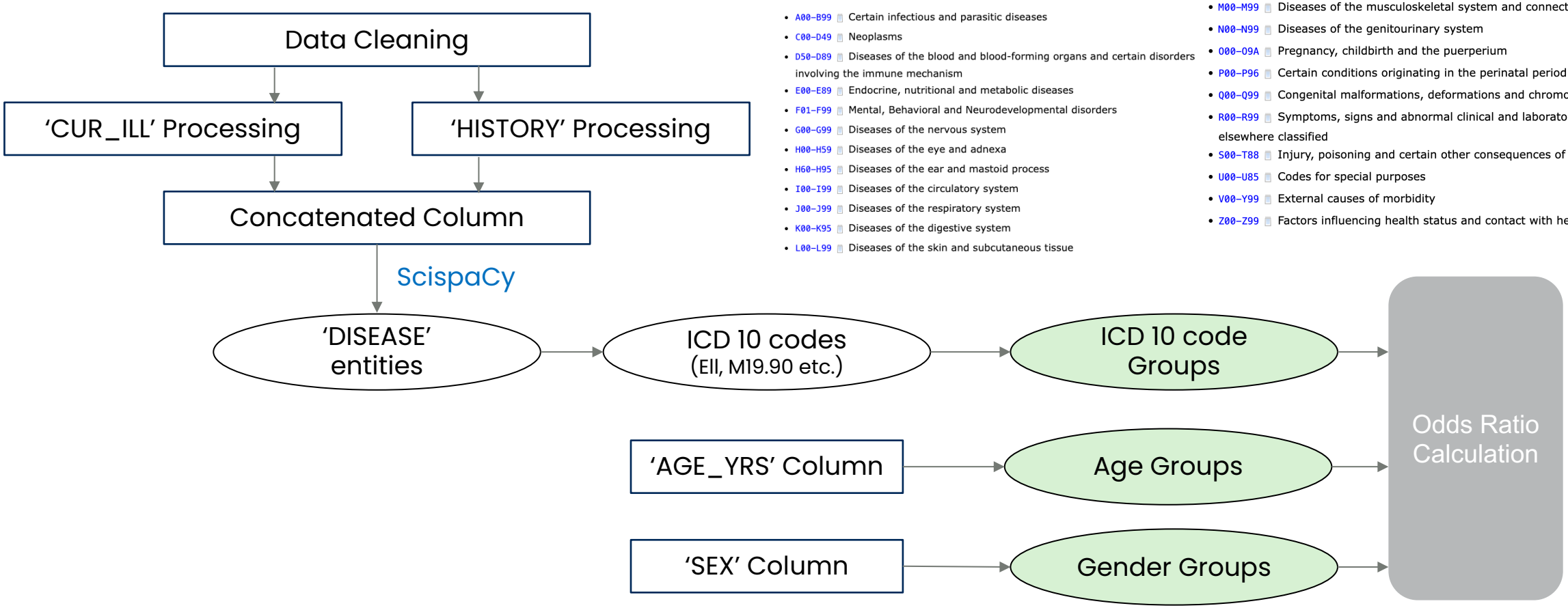
Adverse Event Data from SHINGRIX Clinical Trial

Adverse Reactions	Aged 50-59 Years		Aged 60-69 Years		Aged ≥70 Years	
	SHINGRIX	Placebo[c]	SHINGRIX	Placebo[c]	SHINGRIX	Placebo[c]
Local Adverse Reactions	**n = 1,315 %**	**n = 1,312 %**	**n = 1,311 %**	**n = 1,305 %**	**n = 2,258 %**	**n = 2,263 %**
Pain	88	14	83	11	69	9
Pain, Grade 3[d]	10	1	7	1	4	0.2
Redness	39	1	38	2	38	1
Redness, >100 mm	3	0	3	0	3	0
Swelling	31	1	27	1	23	1
Swelling, >100 mm	1	0	1	0	1	0
General Adverse Reactions	**n = 1,315 %**	**n = 1,312 %**	**n = 1,309 %**	**n = 1,305 %**	**n =2,252 %**	**n = 2,264 %**
Myalgia	57	15	49	11	35	10
Myalgia, Grade 3[e]	9	1	5	1	3	0.4
Fatigue	57	20	46	17	37	14
Fatigue, Grade 3[e]	9	2	5	1	4	1
Headache	51	22	40	16	29	12
Headache, Grade 3[e]	6	2	4	0.2	2	0.4
Shivering	36	7	30	6	20	5
Shivering, Grade 3[e]	7	0.2	5	0.3	2	0.3
Fever	28	3	24	3	14	3
Fever, Grade 3[f]	0.4	0.2	1	0.2	0.1	0.1
GI[g]	24	11	17	9	14	8
GI, Grade 3[e]	2	1	1	1	1	0.4

- *Source: Shingrix prescribing information (FDA)*
- *Based on 2 studies (NCT01165177 & NCT01165229)*

04 SUBGROUP ANALYSIS

PROCESS OVERVIEW



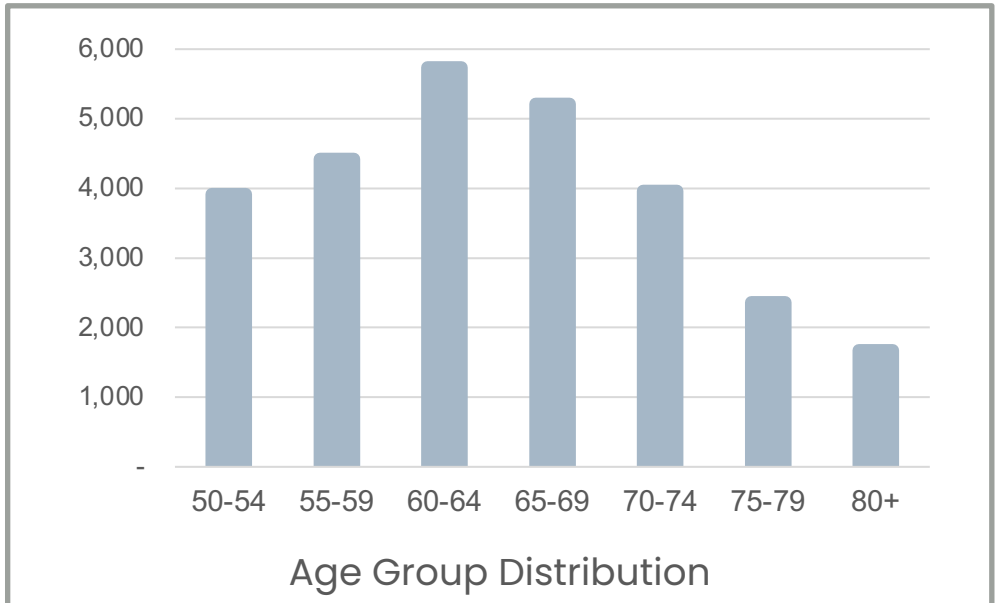
2024 ICD-10-CM Codes

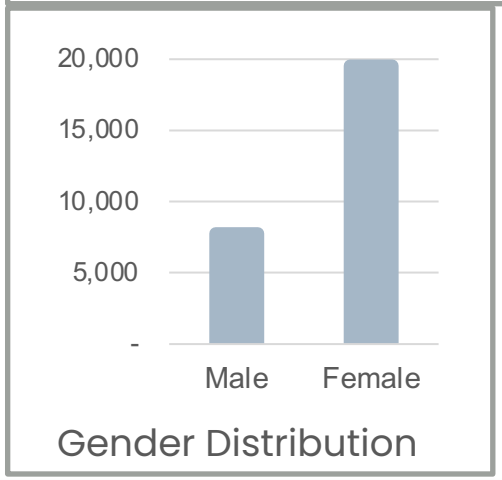
- A00–B99 Certain infectious and parasitic diseases
- C00–D49 Neoplasms
- D50–D89 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- E00–E89 Endocrine, nutritional and metabolic diseases
- F01–F99 Mental, Behavioral and Neurodevelopmental disorders
- G00–G99 Diseases of the nervous system
- H00–H59 Diseases of the eye and adnexa
- H60–H95 Diseases of the ear and mastoid process
- I00–I99 Diseases of the circulatory system
- J00–J99 Diseases of the respiratory system
- K00–K95 Diseases of the digestive system
- L00–L99 Diseases of the skin and subcutaneous tissue
- M00–M99 Diseases of the musculoskeletal system and connective tissue
- N00–N99 Diseases of the genitourinary system
- O00–O9A Pregnancy, childbirth and the puerperium
- P00–P96 Certain conditions originating in the perinatal period
- Q00–Q99 Congenital malformations, deformations and chromosomal abnormalities
- R00–R99 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- S00–T88 Injury, poisoning and certain other consequences of external causes
- U00–U85 Codes for special purposes
- V00–Y99 External causes of morbidity
- Z00–Z99 Factors influencing health status and contact with health services

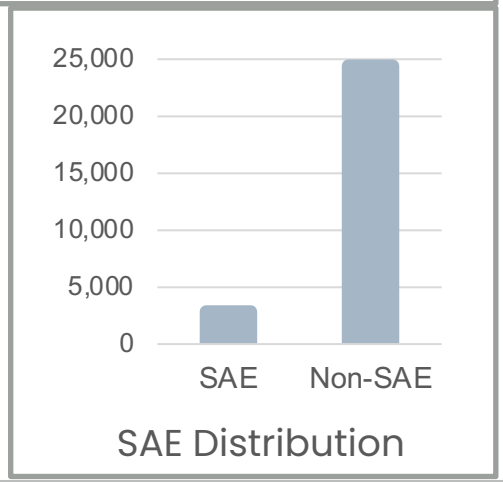
DATA DISTRIBUTION

Conditions Distribution

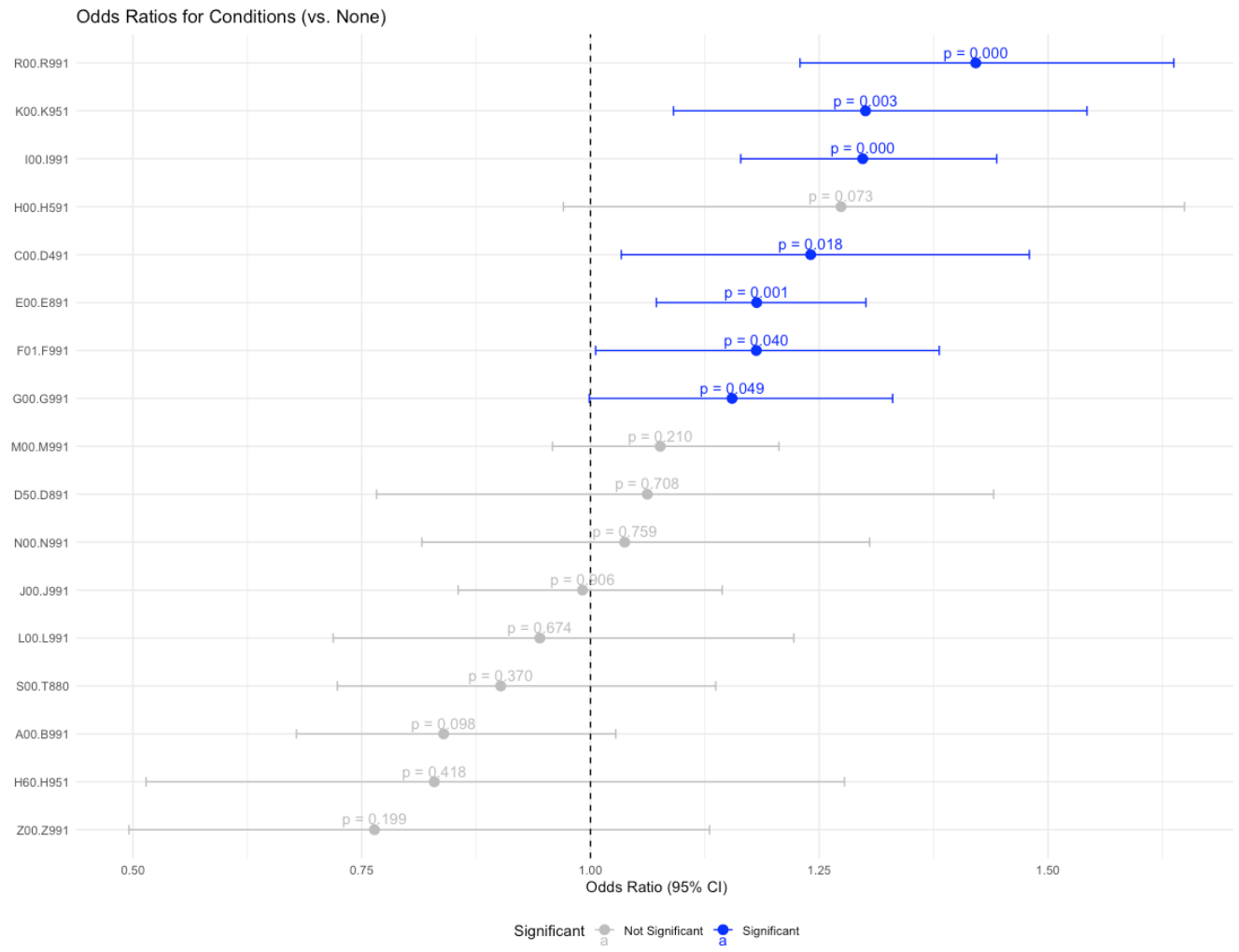
Code	Description (Condition Categories)	Count
A00-B99	Certain infectious and parasitic diseases	1,343
C00-D49	Neoplasms	1,250
D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	400
E00-E89	Endocrine, nutritional and metabolic diseases	6,538
F01-F99	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 musculoskeletal 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 abnormalities	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







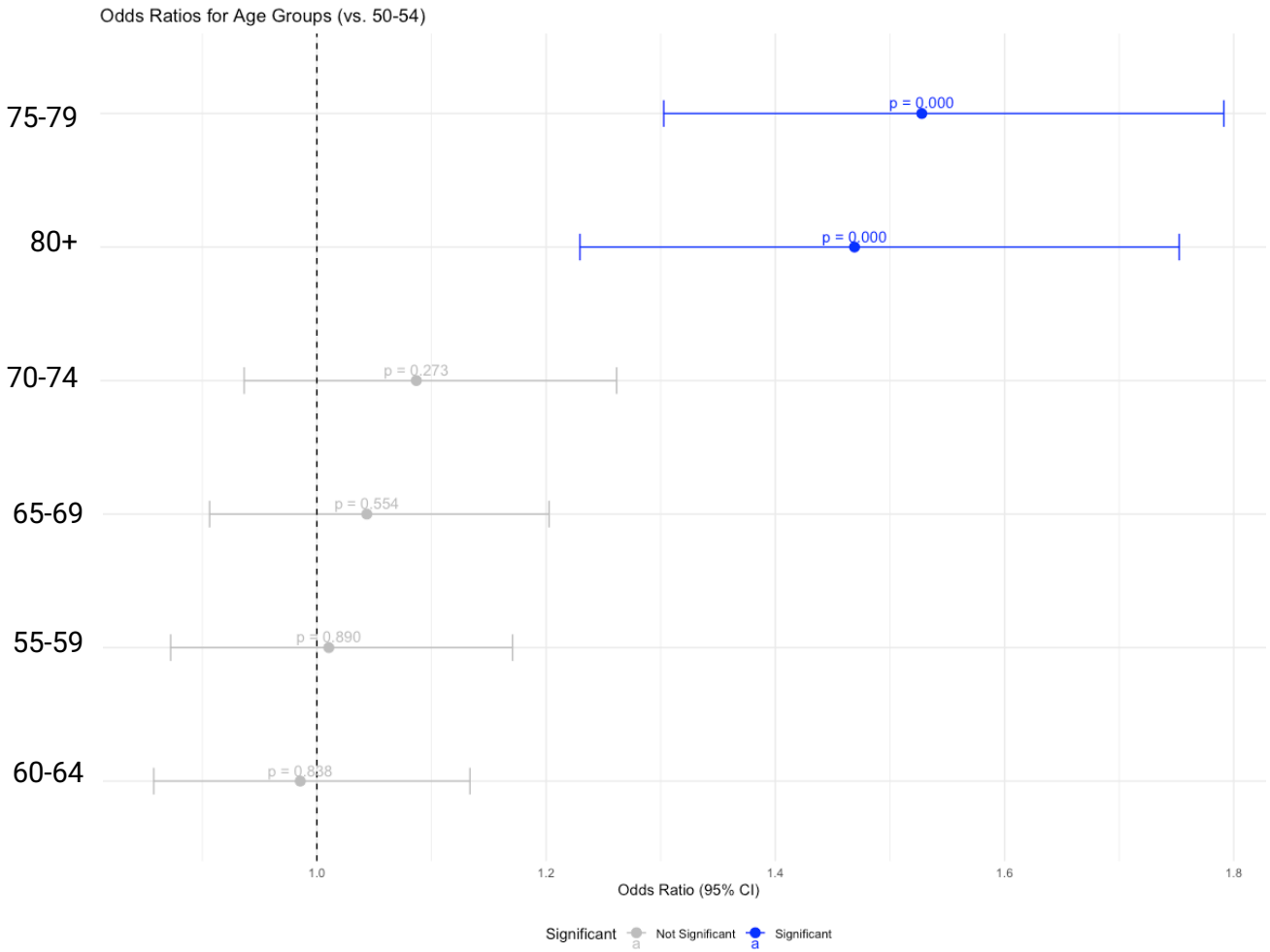
SUBGROUP ANALYSIS RESULTS (VS ‘no condition’)



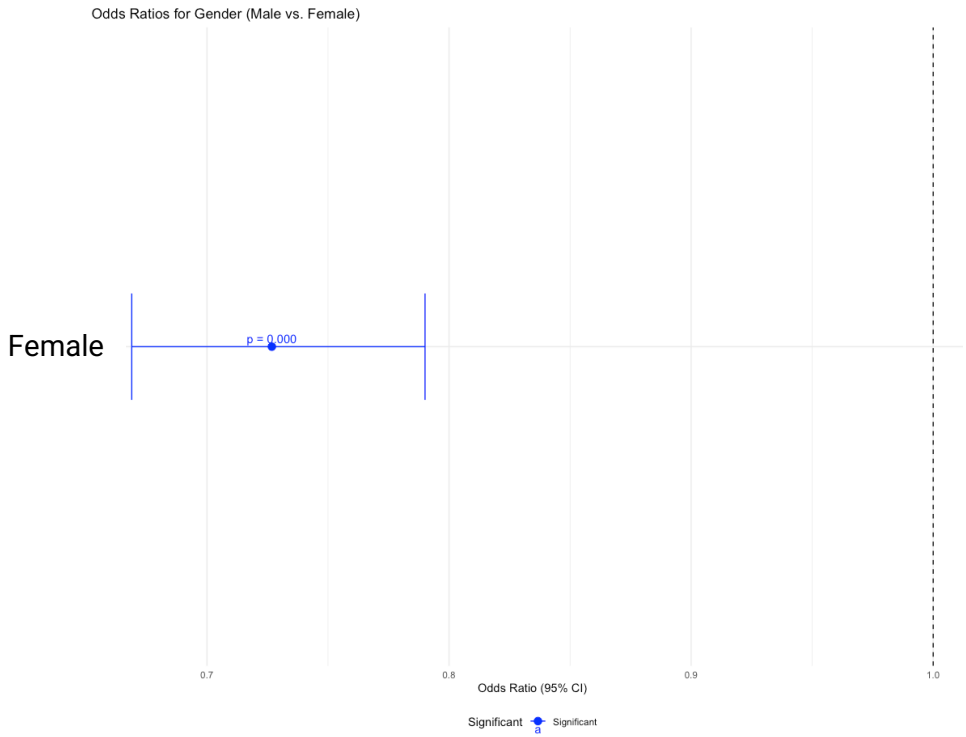
- Individuals with ICD-10 codes R00-R99, K00-K95, I00-I99, C00-D49, E00-E89, F01-F98, and G00-G99 exhibit significantly increased likelihood of SAE compared to those without conditions.
- ICD code groups with lower odds did not show statistically significant results.

R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	Cough, rash, fever
K00-K95	Diseases of the digestive system	Acid reflux, ulcer
I00-I99	Diseases of the circulatory system	Hypertension
C00-D49	Neoplasms	Cancer
E00-E89	Endocrine, nutritional and metabolic diseases	Diabetes
F01-F98	Mental, Behavioral and Neurodevelopmental disorders	Depression, anxiety
G00-G99	Diseases of the nervous system	Migraine

SUBGROUP ANALYSIS RESULTS (VS 50-54, VS male)



- Age groups 70 and older indicated significantly increased odds of SAE occurrence compared to 50-54 age group.
- Female group showed lower odds compared to male group.



05 CONCLUSION & DISCUSSION

DISCUSSION

- BlueBERT + Elastic Net model was superior compared to other models to predict SAE following the onset of AEs using the VAERS dataset with symptom, condition, age, and gender as input features.
- SHAP analysis highlighted the importance of terms such as ‘pain’, ‘fever’, ‘rash’, and ‘headache’ in predicting SAE, which aligns with clinical trial data.
- Certain ICD code groups (hypertension, diabetes, etc.) and older age groups were linked with higher odds of SAE.

LIMITATIONS

- VAERS data – variable quality in reports, symptom description accuracy, and potential bias from post-SAE reporting.
  → Subgroup analysis (R group with AE symptoms identified as highest odds ratio)
  → Possible SAE information leakage in symptom text

CONCLUSION & FUTURE WORK

- This study enables the prediction of SAE progression at the onset of AEs, allowing for proactive patient care and intervention.
- Statistical analysis identified specific medical conditions, age groups, and gender as risk factors for SAE development, facilitating personalized care decisions.
- Further model refinement (parameters, different models) may improve prediction performance.
- Including additional features (co-administered vaccines, concomitant medications etc.) may provide a more comprehensive understanding.

THANK YOU!

REFERENCES

1. The Washington Post. (2023, May 17). *Vaccine hesitancy.* https://www.washingtonpost.com/opinions/2023/05/17/poll-vaccine-hesitancy-doctors/
2. Pharmacy Times. (n.d.). *Survey identifies factors in herpes zoster vaccine hesitancy among older adults.* https://www.pharmacytimes.com/view/survey-identifies-factors-in-herpes-zoster-vaccine-hesitancy-among-older-adults
3. Centers for Disease Control and Prevention. *Shingles (Herpes Zoster).* https://www.cdc.gov/shingles/about/index.html
4. ID Care. *Shingles: All You Need to Know.* https://idcare.com/blog/shingles-all-you-need-to-know/
5. Centers for Disease Control and Prevention. Vaccine Information Statement (VIS): *Shingles (Recombinant) VIS.* https://www.cdc.gov/vaccines/hcp/vis/vis-statements/shingles-recombinant.html
6. Szabo, A., & McMeekin, S. (2023). *Herpes zoster and postherpetic neuralgia*. *Drugs & Aging*, 39(11), 989–1001.
7. 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.
8. 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-andpandemic-has-made-it-worse.
9. (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.
10. (WHO), W.H.O., *The conceptual framework for the international classification for patient safety.* 2009.
11. Turing. *Guide on Word Embeddings in NLP.* July 24, 2024]; Available from: https://www.turing.com/kb/guide-on-word-embeddings-in-nlp.
12. Ramos, J.E. *Using TF-IDF to Determine Word Relevance in Document Queries.* 2003.
13. 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.
14. 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.
15. 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.
16. scikit-learn. *StandardScaler.* July 24, 2024]; Available from: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.
17. Sperandei, S., *Understanding logistic regression analysis.* Biochemia Medica, 2014. 24(1): p. 12-18.
18. 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.
19. 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.
20. Lundberg, S.M., & Lee, S. I. . *SHAP (SHapley Additive exPlanations).* 2017; Available from: https://github.com/shap/shap.
21. 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.
22. 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.
23. ICD10Data.com*, ICD-10-CM Codes.*
24. (FDA), *F.a.D.A. Shingrix Prescribing Information.* 2023 [cited July 25, 2024; Available from: https://www.fda.gov/media/108597/download.