Identifying Vaccine Hesitancy Patterns – A Data Driven Approach to Public Health Intervention

Table of Contents

01 Motivation & Business Problem

02 Data Understanding

03 Analytical Process

04 Outputs & Findings

05 Clustering Analysis

06 Cost Analysis

07 Key Insights & Recommendations

08 Future Improvements

Motivation & Business Problem

Vaccination is one of the most significant advancements in medical science. Yet there are preventable diseases that continue to persist. Why?

Vaccination Refusal vs. Vaccination Hesitancy

The Problem: Public health campaigns risk inefficient resource allocation and limited impact on vaccination uptake.

Our Data

Data from the CDC’s National Center for Health Statistics (NCHS).

2009-2010 H1N1 dataset containing over 26,000 records.

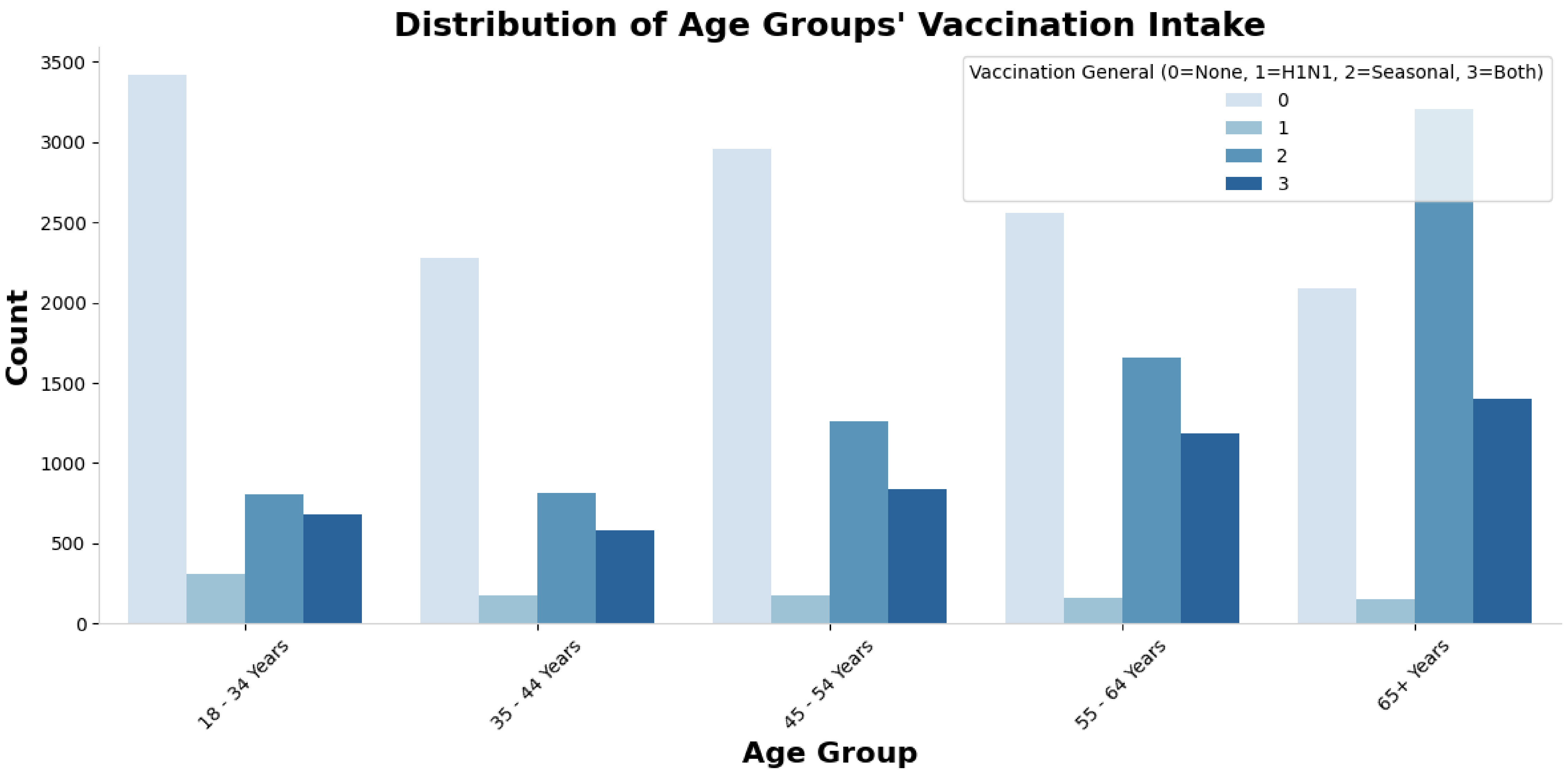
Survey Data of all categorical variables.

Topic include people’s hygienic behaviors, avoidance behaviors, perceptions of risks, healthcare outreach, and demographic information.

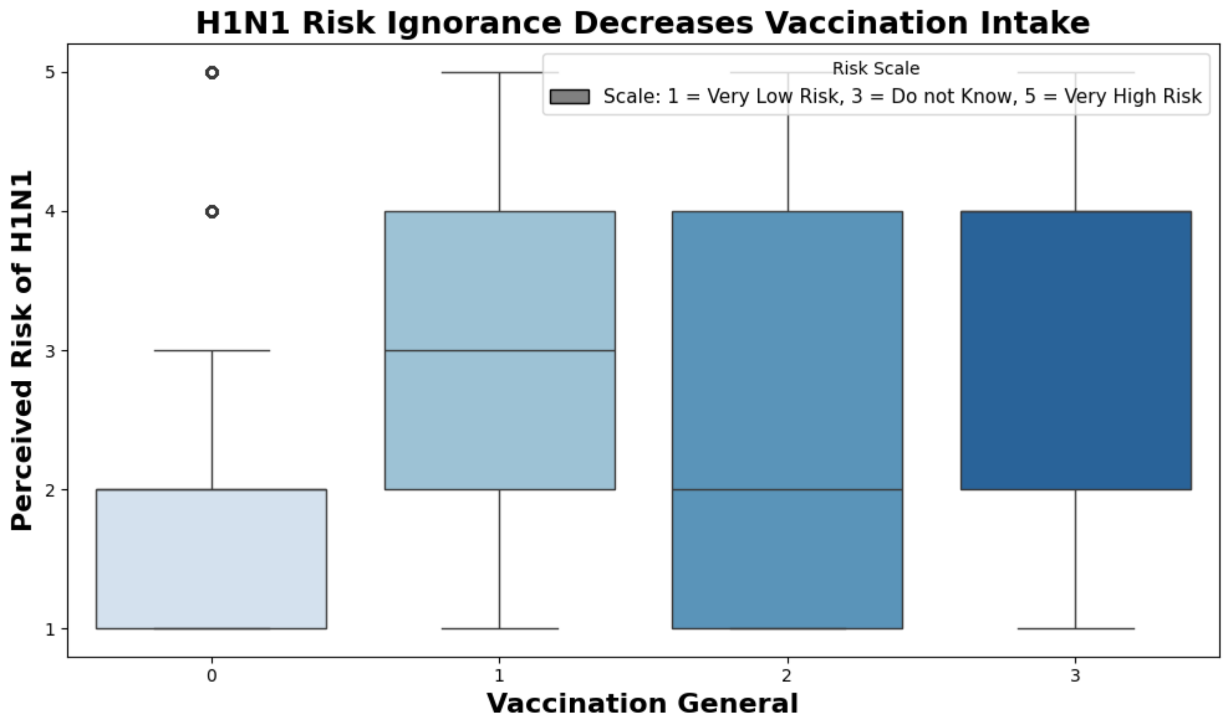
Variables Dictionary

Variable Name	Variable Dictionary Name	Description
Vaccination Status	**Vaccination_General**	**Target variable whether individuals got either h1n1 or seasonal vaccine or not (binary: 0 = No, 1 = h1n1 only, 2 = seasonal only, 3 = both)**
Doctor H1N1 Vaccine Recommendation	doctor_recc_h1n1	H1N1 flu vaccine was recommended by doctor (binary)
Doctor Flu Vaccine Recommendation	doctor_recc_seasonal	Seasonal flu vaccine was recommended by doctor (binary)
Individual Has Health Insurance	health_insurance	Has health insurance (binary)
Opinion on Contracting H1N1	opinion_h1n1_risk	Opinion on risk of getting H1N1 without vaccine (1 = Very low, 3 = Don't know, 5 = Very high)
Opinion on Effectiveness of Seasonal Vaccine	opinion_seas_vacc_effective	Opinion on seasonal flu vaccine effectiveness (1 = Not at all, 3 = Don't know, 5 = Very effective)
Opinion on Contracting Seasonal Flu	opinion_seas_risk	Opinion on risk of getting seasonal flu without vaccine (1 = Very low, 3 = Don't know, 5 = Very high)
Individual Age Group	age_group	Age group of respondent
Employment Status	employment_industry	Industry respondent is employed in (21 coded values)
H1N1 Level of Concern	h1n1_concern	Level of concern about H1N1 flu (0 = Not at all, 1 = Not very, 2 = Somewhat, 3 = Very)
H1N1 Level of Knowledge	h1n1_knowledge	Level of knowledge about H1N1 flu (0 = No knowledge, 1 = A little, 2 = A lot)
Individual Has Chronic Medical Conditions	chronic_med_condition	Has chronic medical conditions (e.g., asthma, diabetes, heart/kidney condition) (binary)
Race	race	Race of respondent
Sex	sex	Sex of respondent
Household Income	income_poverty	Household income relative to 2008 Census poverty thresholds. (Above or Below $75,000)

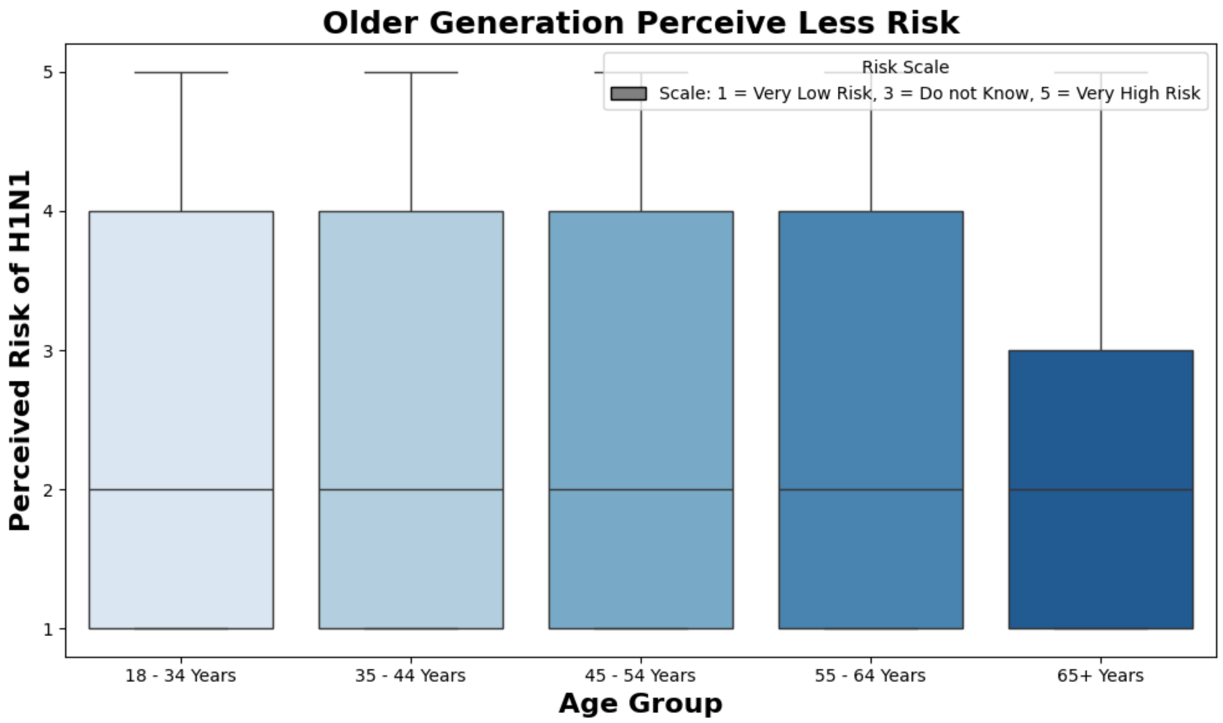
Persistent Hesitancy is a Cross Generational Challenge



Diseases Risk Ignorance Reduces Vaccination Intake

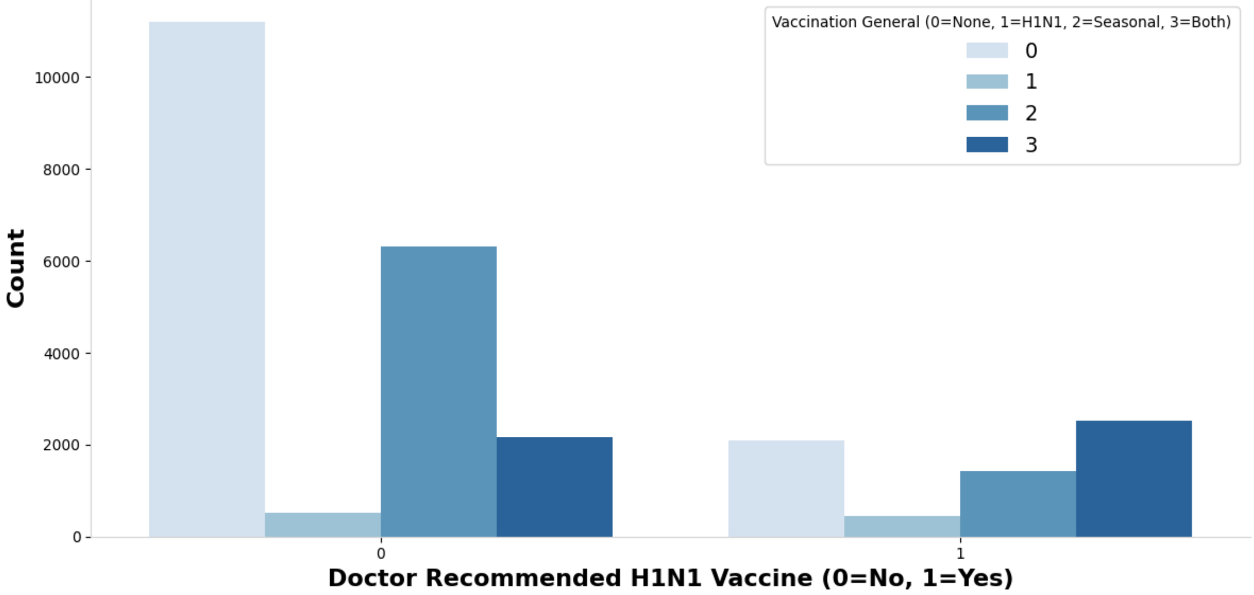


H1N1 Risk Ignorance Decreases Vaccination Intake

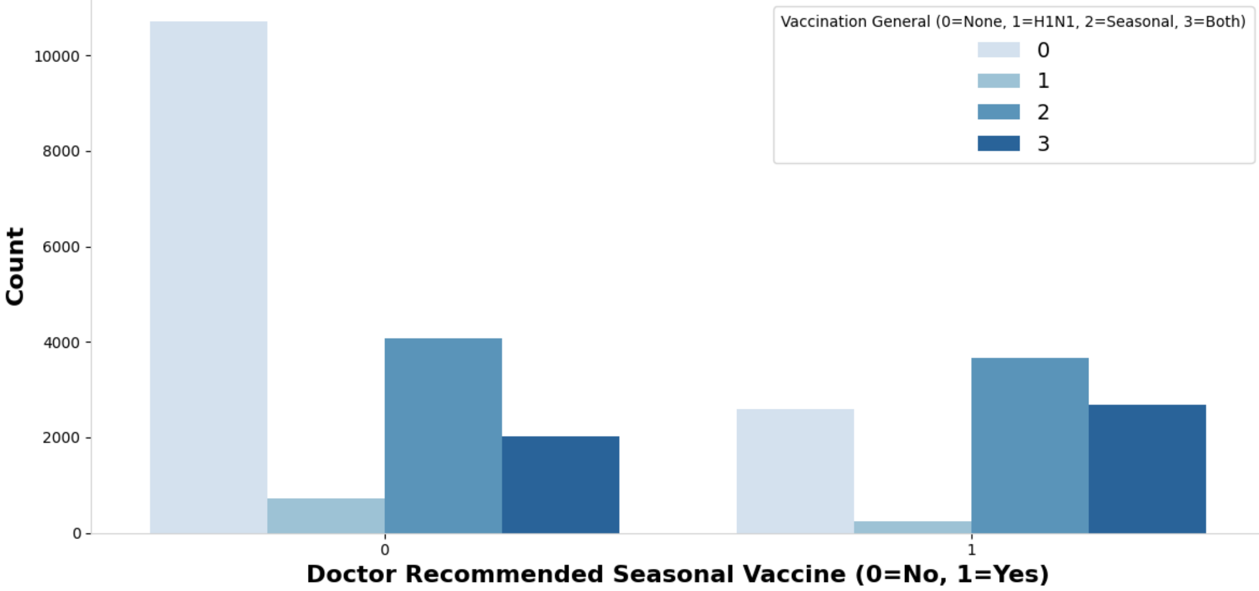


Older Generation Perceive Less Risk

Doctors Opinion Matters

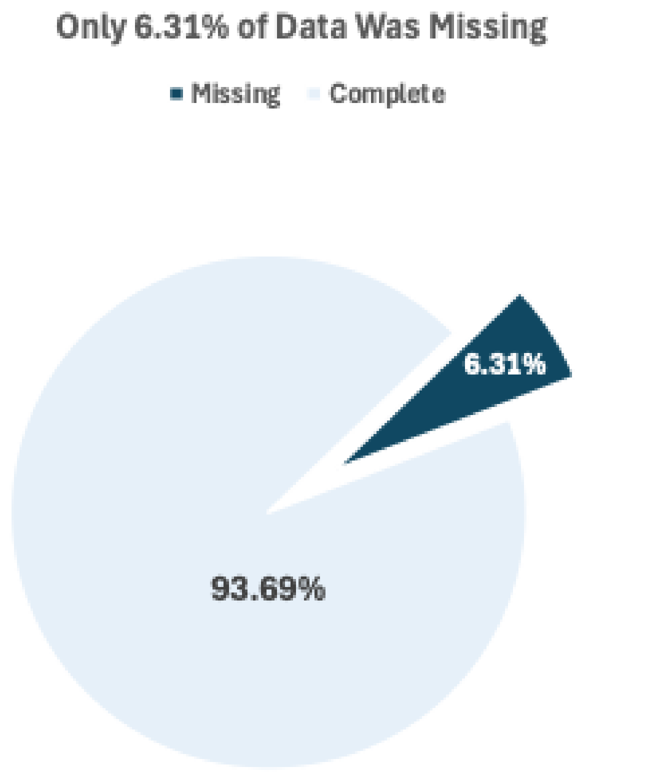


Effect of Doctor Recommendation on H1N1 Vaccination Intake

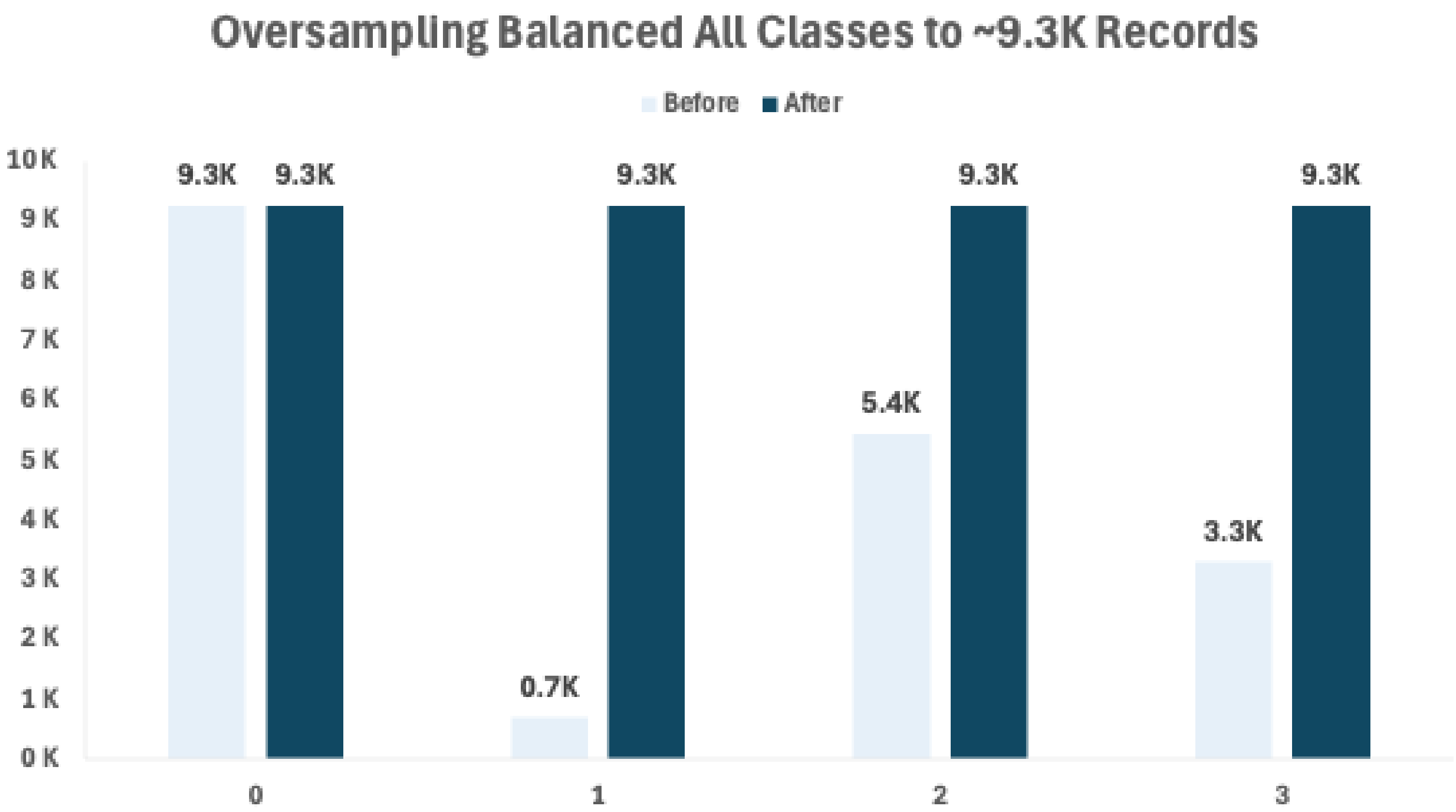


Effect of Doctor Recommendation on Seasonal Vaccination Intake

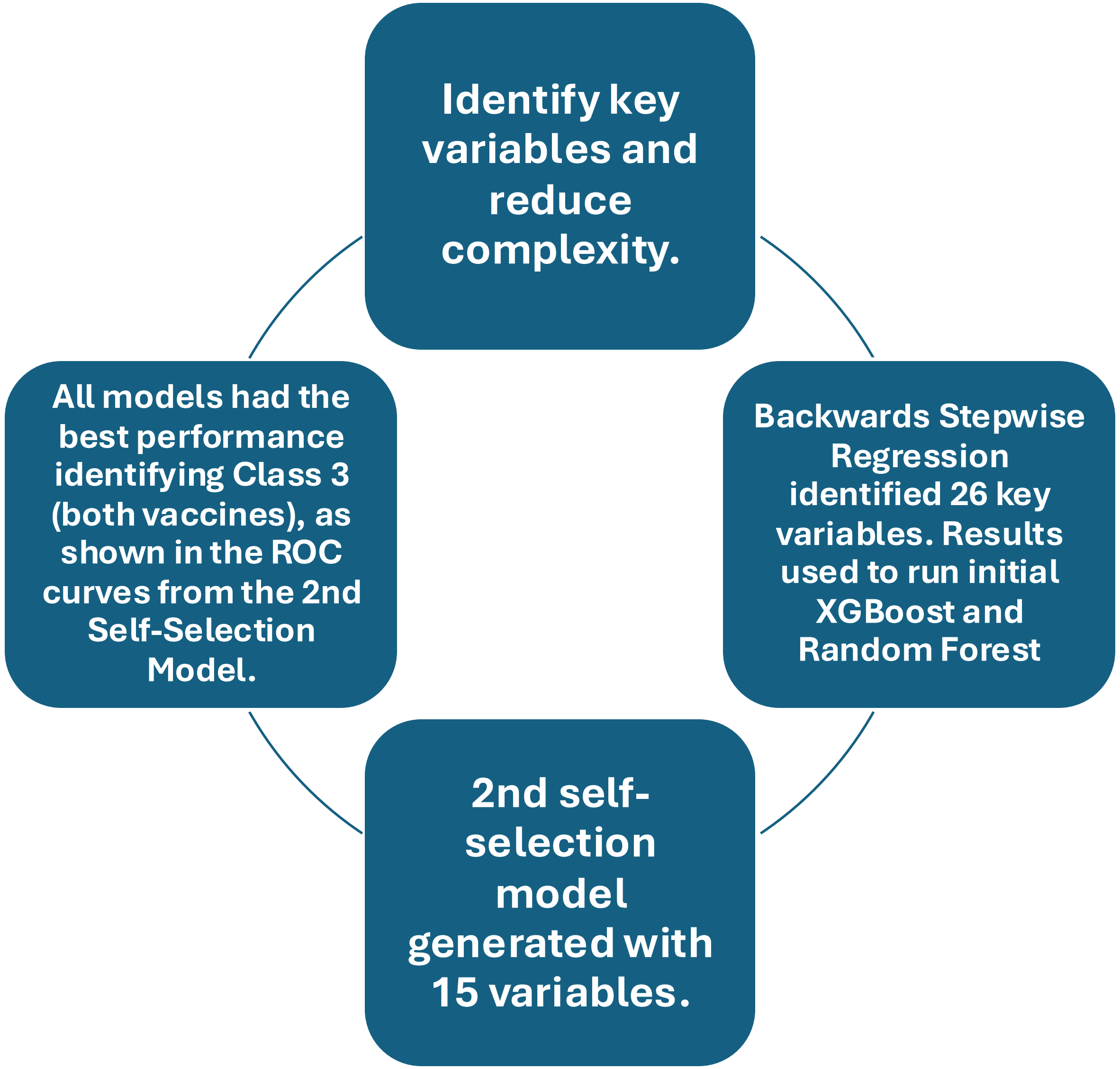
Analytical Process



Missing values were imputed using random sampling and equal reassignment across existing classes.



Multinomial Logistic Regression



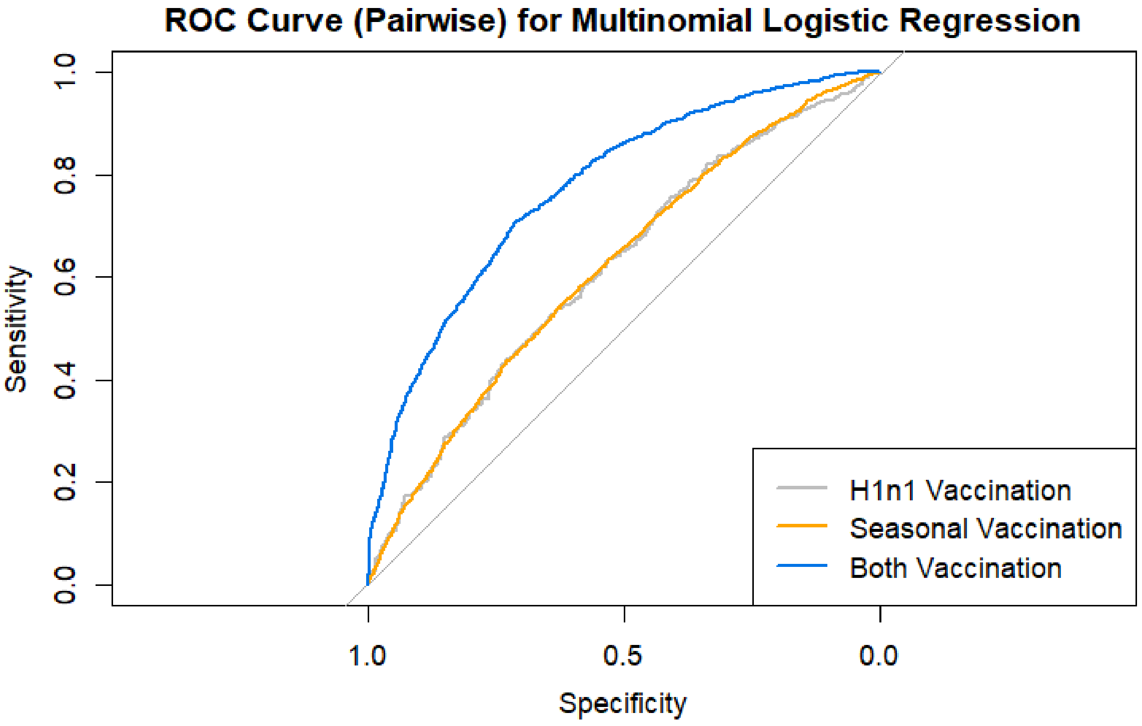
Identify key variables and reduce complexity.

Backwards Stepwise Regression identified 26 key variables. Results used to run initial XGBoost and Random Forest

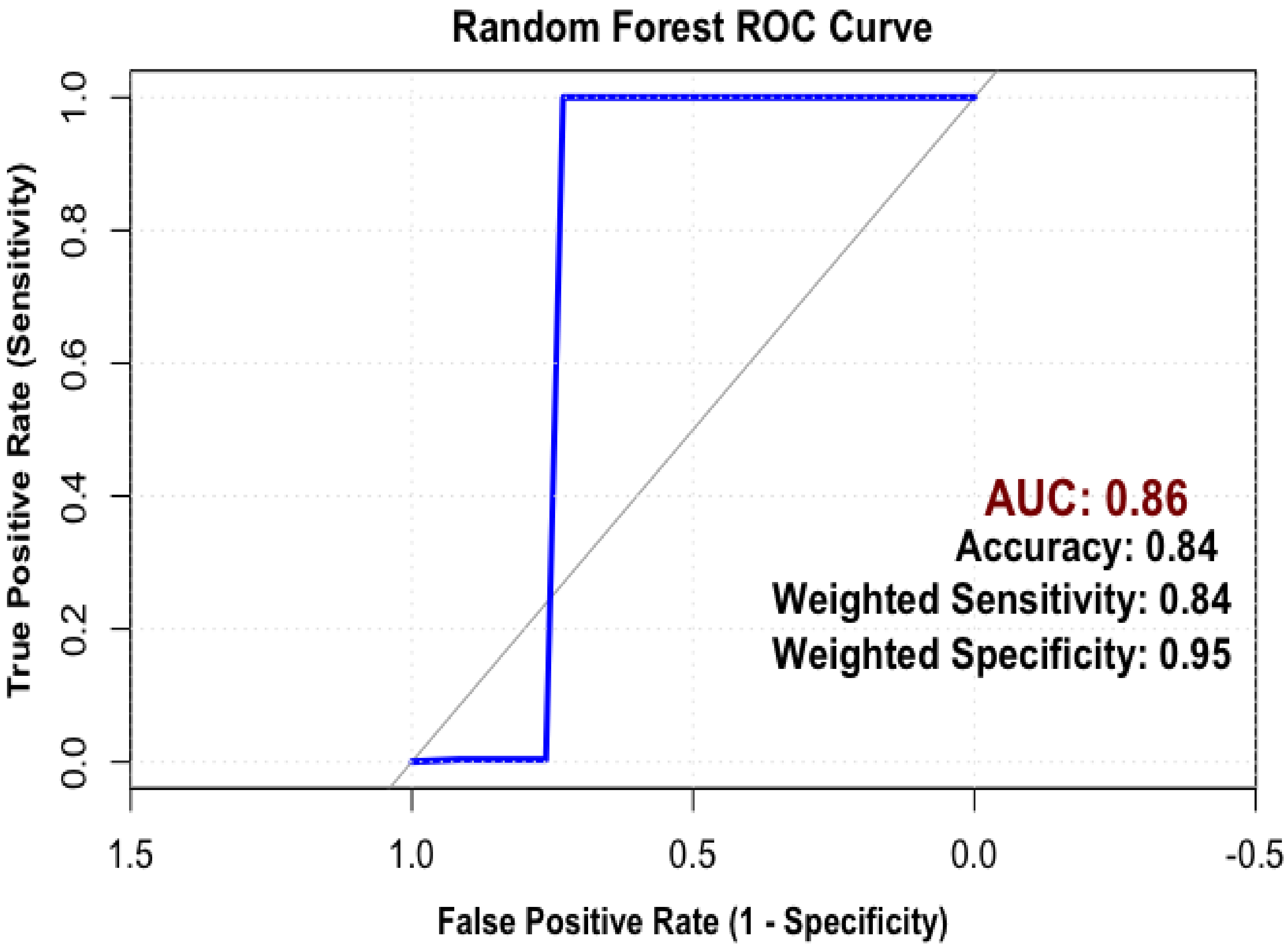
2nd self-selection model generated with 15 variables.

All models had the best performance identifying Class 3 (both vaccines), as shown in the ROC curves from the 2nd Self-Selection Model.

	Final Multiclass Logistic Model		
	Self-Selection	Backwards Stepwise	2nd Self-Selection
Accuracy	58.23%	59.15%	43.45%
Multiclass AUC	79.78%	79.75%	66.82%
Sensitivity	55.42%	56.09%	40.08%
Specificity	86.01%	86.17%	80.66%

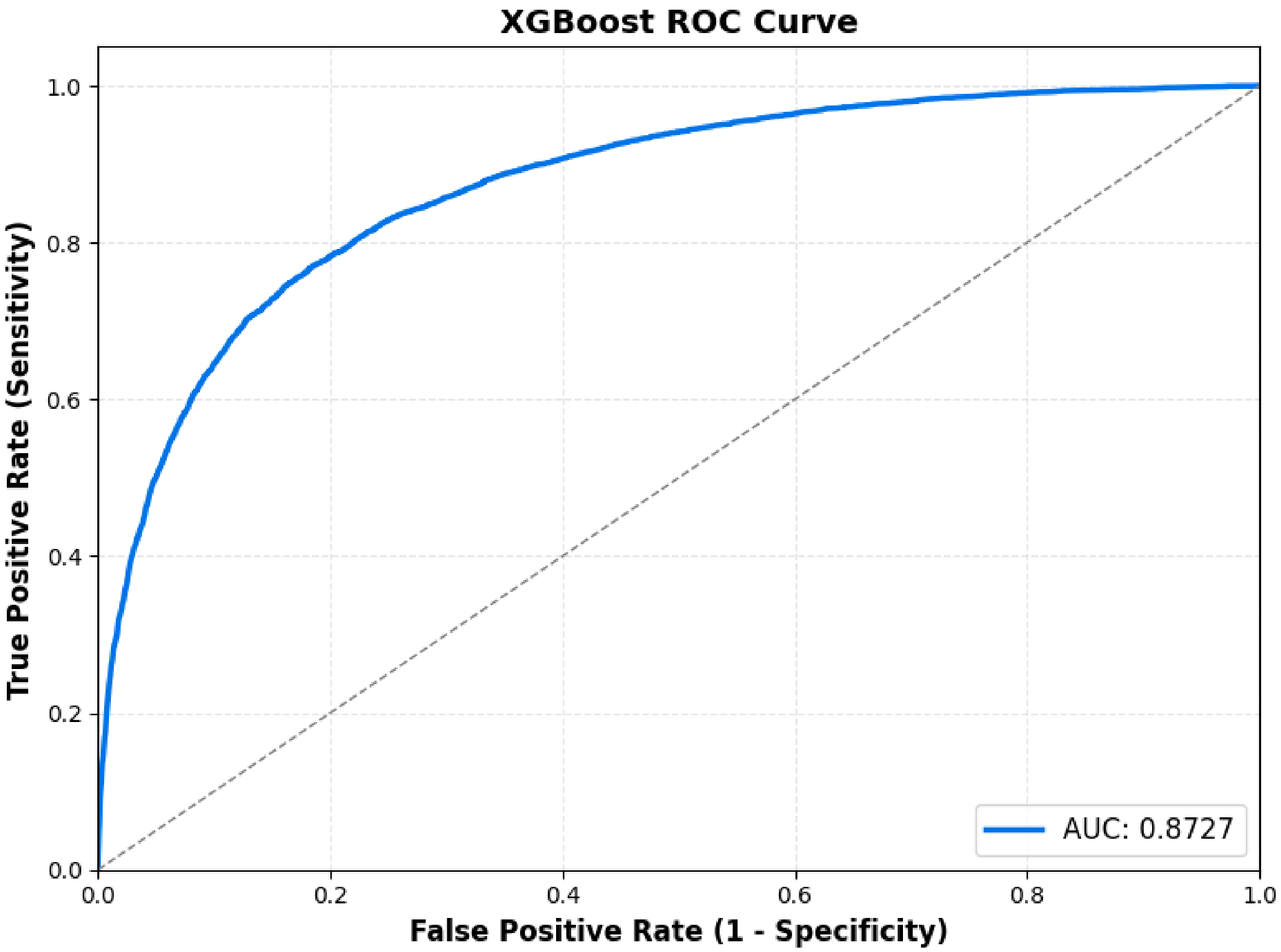


Random Forest Model



The steep, square-shaped ROC curve results from the model's extremely high sensitivity and specificity for Class 1, indicating near-perfect discrimination and highly confident probability estimates for that class.

XGBoost Model



- The model achieved an AUC of 0.8727, indicating strong overall ability to classify individuals by their vaccination status.

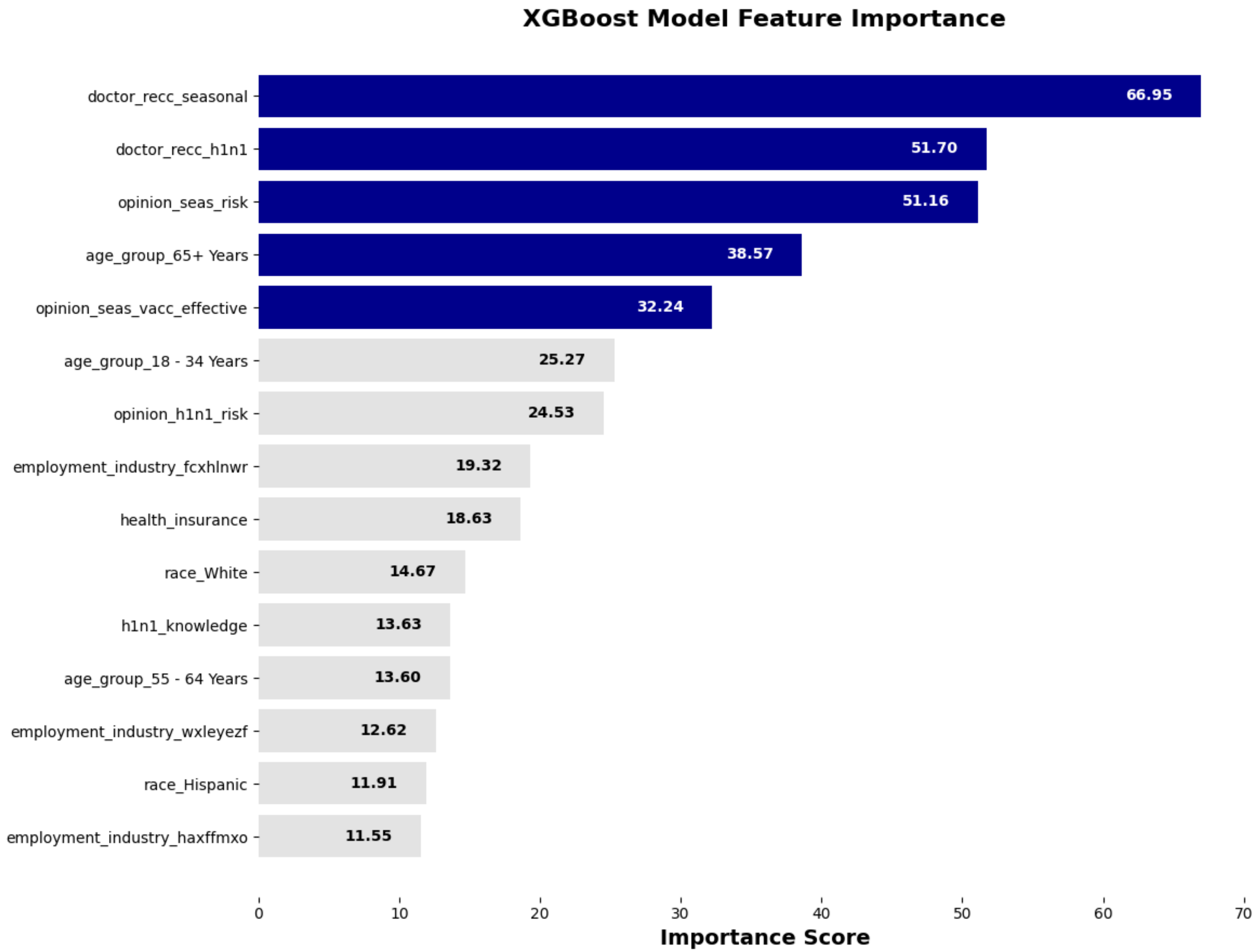
- GridSearchCV was used for hyperparameter tuning, allowing the model to find the optimal combination of settings for improved generalization and balanced performance.

Model Performance Comparison

Model	Accuracy	AUC	Sensitivity	Specificity
Baseline	50%			
Multiclass Logistic	43.45%	66.82%	40.08%	80.66%
Random Forest	84.22%	85.62%	84.22%	94.78%
XGBoost	**66.21%**	**87.27%**	**63.74%**	**87.86%**

XGBoost was chosen for its ability to reduce bias by reliably distinguishing between unvaccinated (sensitivity) and vaccinated (specificity) individuals.

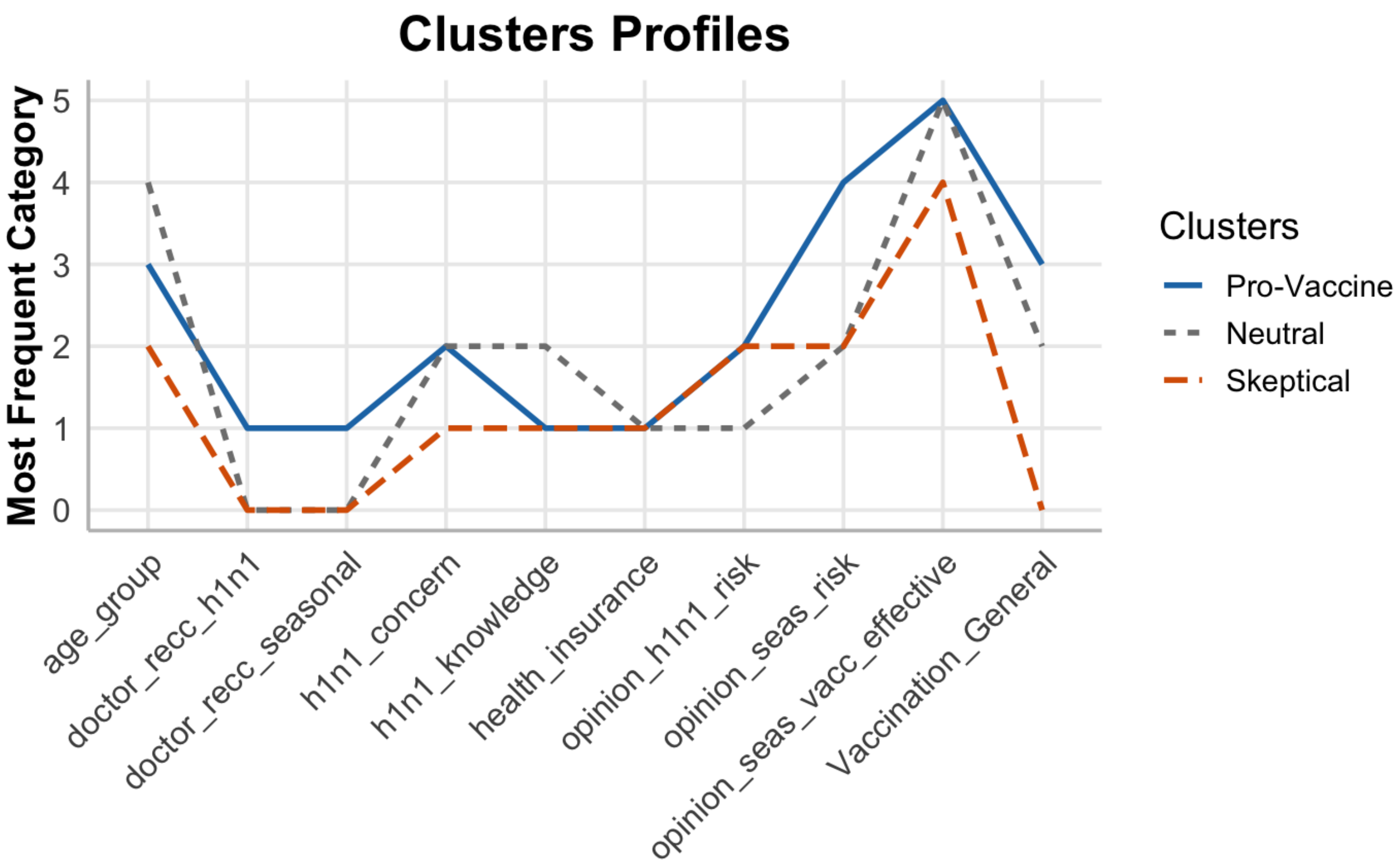
XGBoost Model

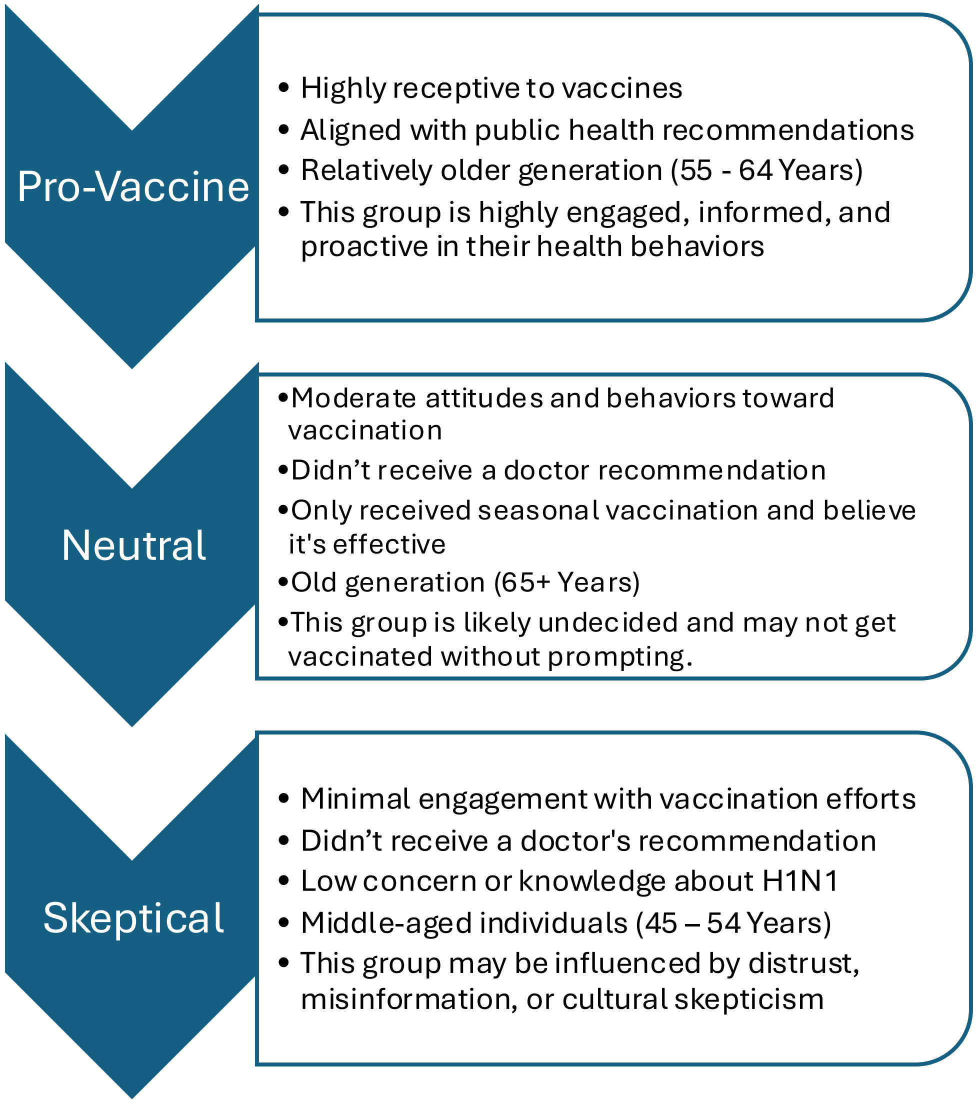


- **Doctor recommendations**, **risk perception**, and **age** emerged as the top drivers of vaccine behavior.

- These results align with public health findings on the importance of trust and risk awareness in vaccine uptake.

Clustering Analysis





Pro-Vaccine

- Highly receptive to vaccines
- Aligned with public health recommendations
- Relatively older generation (55 - 64 Years)
- This group is highly engaged, informed, and proactive in their health behaviors

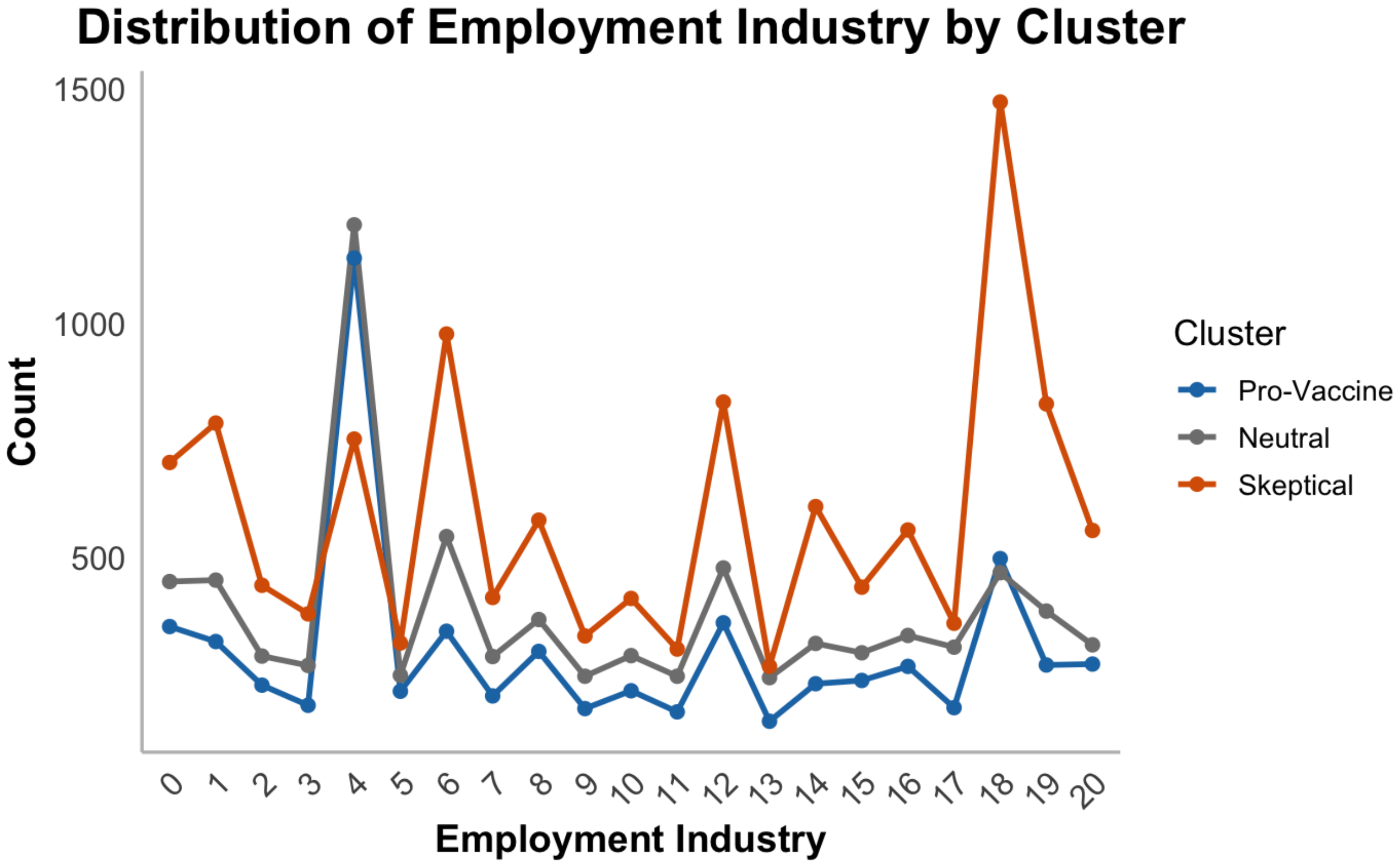
Neutral

- Moderate attitudes and behaviors toward vaccination
- Didn’t receive a doctor recommendation
- Only received seasonal vaccination and believe it's effective
- Old generation (65+ Years)
- This group is likely undecided and may not get vaccinated without prompting.

Skeptical

- Minimal engagement with vaccination efforts
- Didn’t receive a doctor's recommendation
- Low concern or knowledge about H1N1
- Middle-aged individuals (45 – 54 Years)
- This group may be influenced by distrust, misinformation, or cultural skepticism

Clustering Analysis – Employment Industry

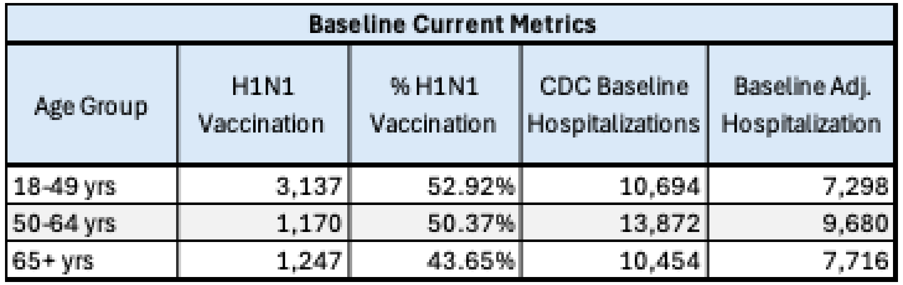


The Industry numbering comes from encrypted data from the CDC, and they can re-identify it.

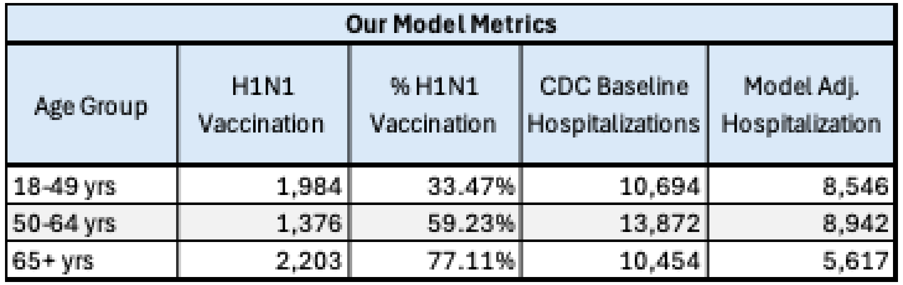
Skeptical: Focused in industry 18

Neutral & Pro-Vaccine: Focused in industry 4

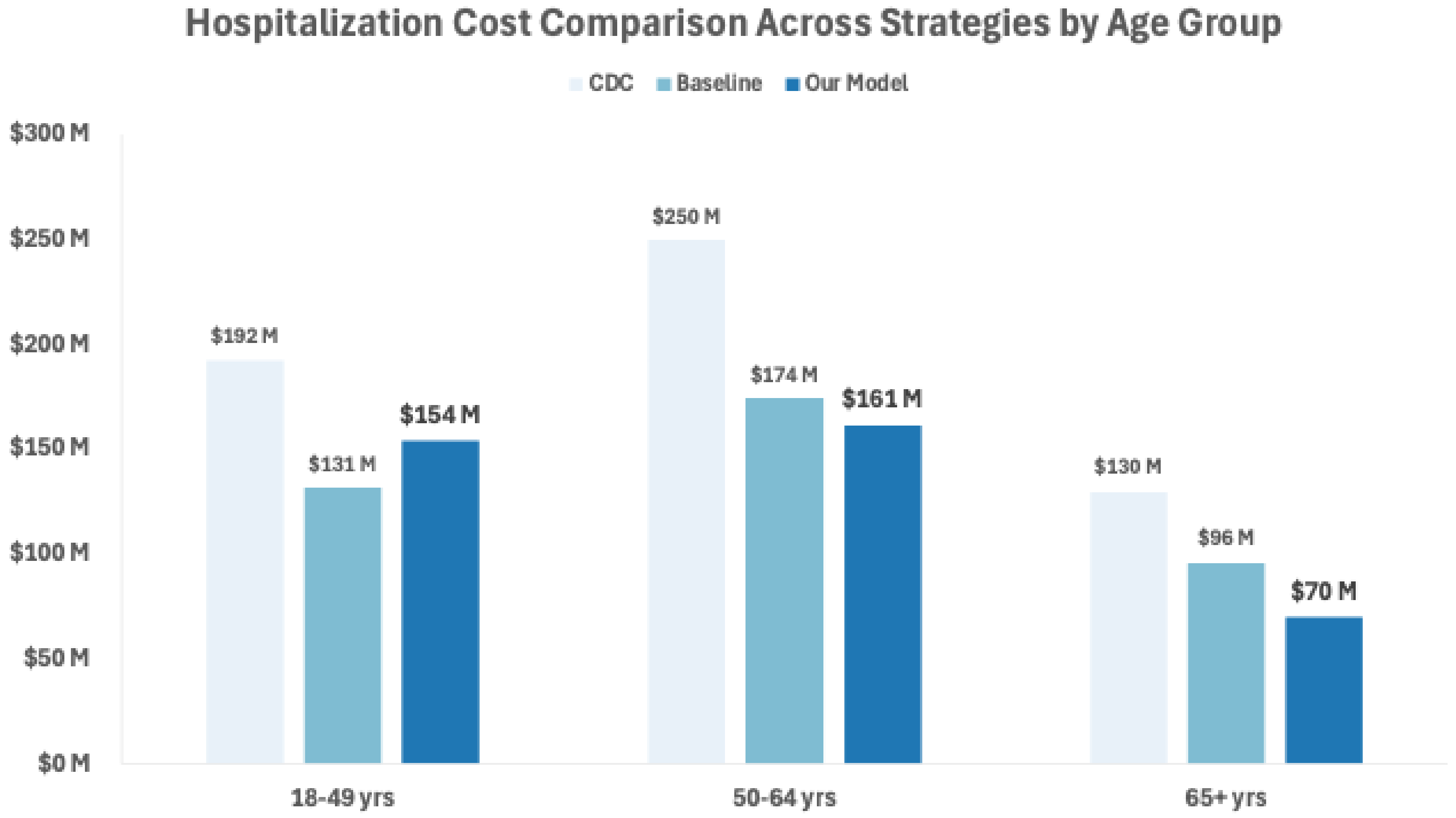
Cost Analysis



Baseline Current Metrics				
Age Group	H1N1 Vaccination	% H1N1 Vaccination	CDC Baseline Hospitalizations	Baseline Adj. Hospitalization
18-49 yrs	3,137	52.92%	10,694	7,298
50-64 yrs	1,170	50.37%	13,872	9,680
65+ yrs	1,247	43.65%	10,454	7,716



Our Model Metrics				
Age Group	H1N1 Vaccination	% H1N1 Vaccination	CDC Baseline Hospitalizations	Model Adj. Hospitalization
18-49 yrs	1,984	33.47%	10,694	8,546
50-64 yrs	1,376	59.23%	13,872	8,942
65+ yrs	2,203	77.11%	10,454	5,617



Cost Analysis

Model Costs	
	Costs
Baseline Costs	$ 401 M
Our Method	**$ 384 M**

Doctor Recommendations	
Age Group	Cost Reduction
One-on-One Counseling	$ 210 M
Strong Provider Recommendation	$ 206 M

Addressing Concern/Knowledge	
Age Group	Cost Reduction
Reminder/Recall Systems	$ 208 M
School-Based Programs	$ 196 M
Multicomponent Education	$ 217 M

Initial Implementation leads to $17 million in cost reduction, compared to current costing methods.

Long-term: Targeted Campaigns related to doctor recommendations, addressing disease knowledge and risk, and education programs.

Education, Providers, and Financial based programs can be utilized.

Targeted Campaigns around key problems would save an addition $196 to $217 million, depending on program.

Key Insights & Recommendations

Many people believe in the effectiveness of the seasonal flu vaccine.

Neutral and Skeptical cluster have low perception of risk of H1N1.

Doctor Recommendation is a big factor in influencing vaccination status.

Focus on Methodology: The value lies in our methodology, providing framework for future endemic or pandemic events.

Tailored Interventions Matter: Vaccine decisions are nuanced; a "one-size-fits-all" approach is ineffective. Precision-targeted outreach addresses hesitancy more effectively.

Health Impact: Increases herd immunity, reduces the strain on Hospitals and allocate resources toward other sectors of healthcare.

Future Improvements

Do younger individuals (18-21) have a similar trend for feature importance and clustering?

Are there different factors that influence the younger demographics that were not taken in account in this analysis?

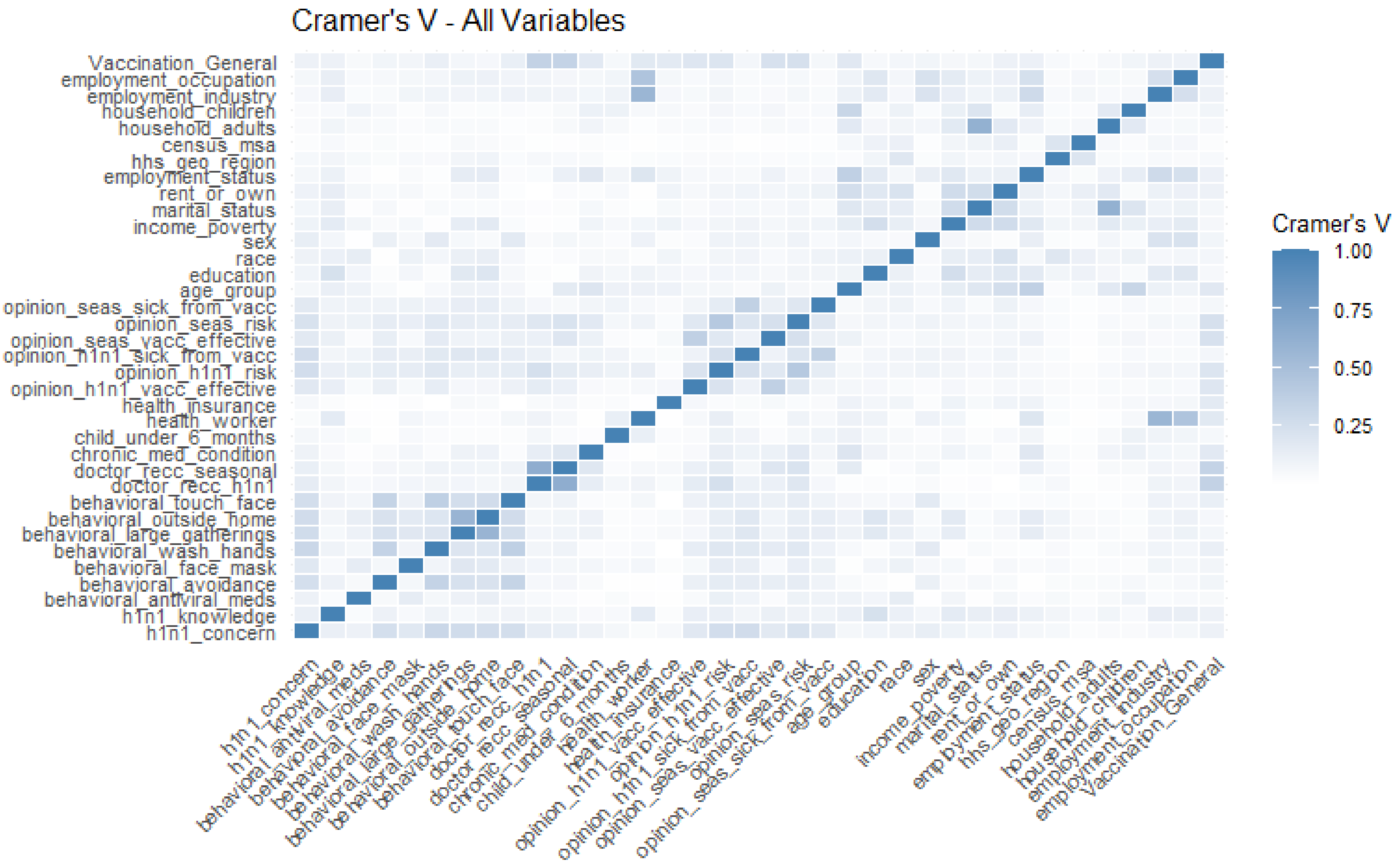
What are the other costs associated with targeted intervention campaigns?

Thank you! Any questions?

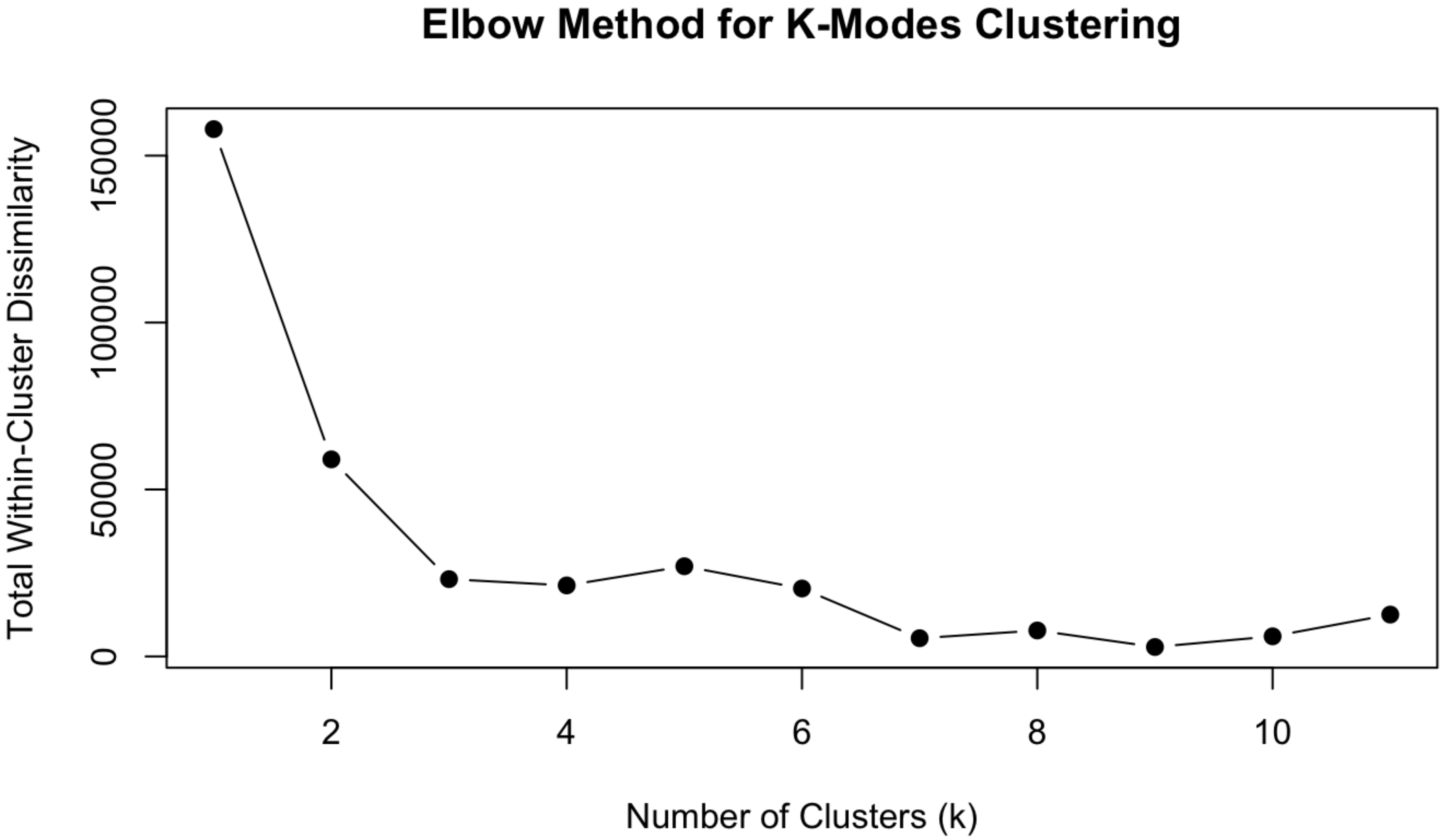
Appendix

- Correlation Matrix
- Number of Clusters Selection Method
- Random Forest Final Model Feature Importance
- All Cost Reductions
- H1N1 & Seasonal Effectiveness By Age Group
- Risk Perception of H1N1 by Age Group
- Vaccination Status by Health Insurance Status
- Vaccination Status by Sex
- Health Insurance By Age Group
- H1N1 Doctor Recommendation By Age Group
- Random Forest Feature Importance Used for Final Variable Selection

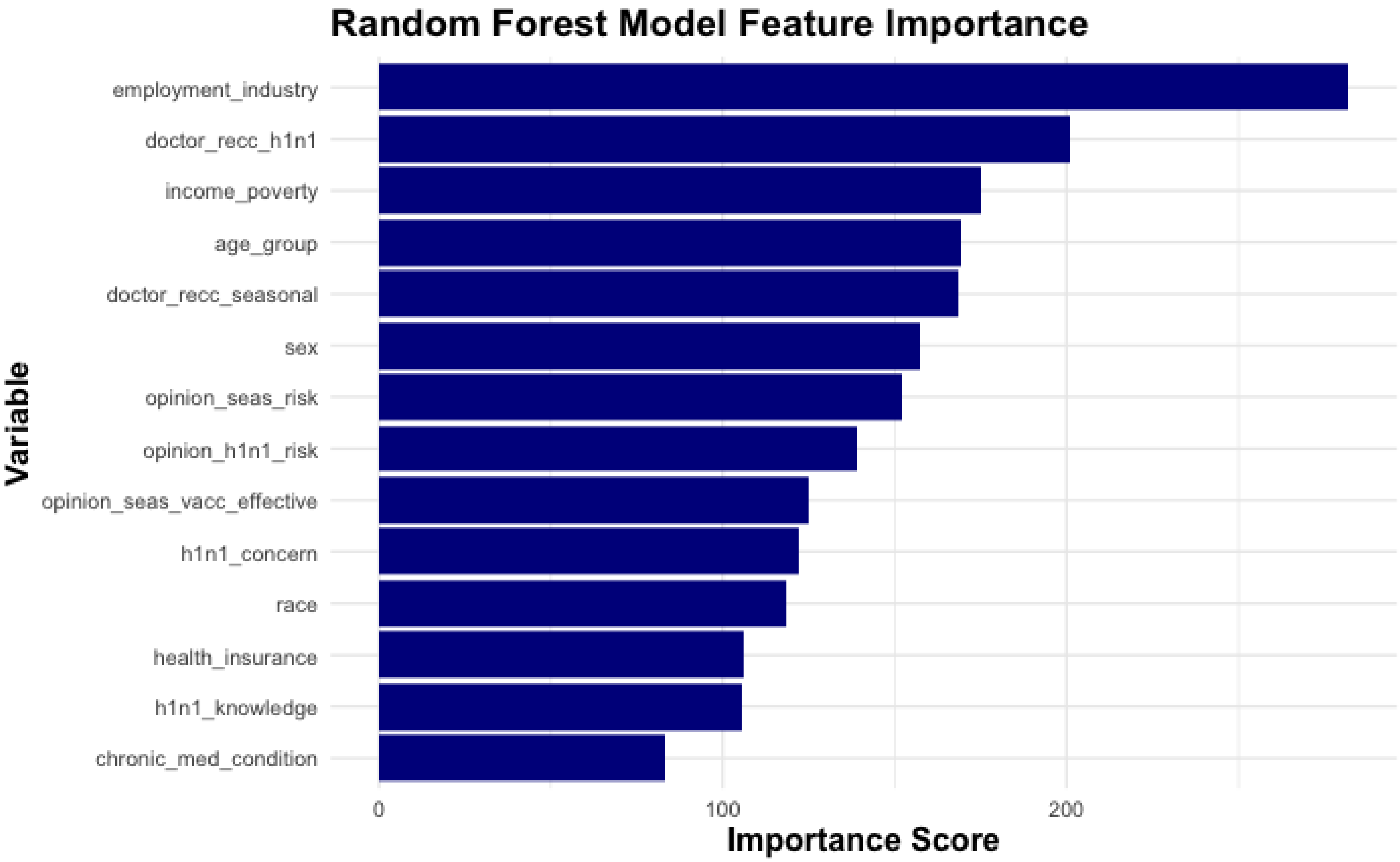
Correlation Matrix



Number of Clusters Selection Method



Random Forest Final Model Feature Importance



All Cost Reductions

Model Costs

	Costs
Baseline Costs	$ 401 M
Our Method	**$ 384 M**

Addressing Concern/Knowledge

Age Group	Cost Reduction
Reminder/Recall Systems	$ 208 M
School-Based Programs	$ 196 M
Multicomponent Education	$ 217 M

Doctor Recommendations

Age Group	Cost Reduction
One-on-One Counseling	$ 210 M
Strong Provider Recommendation	$ 206 M

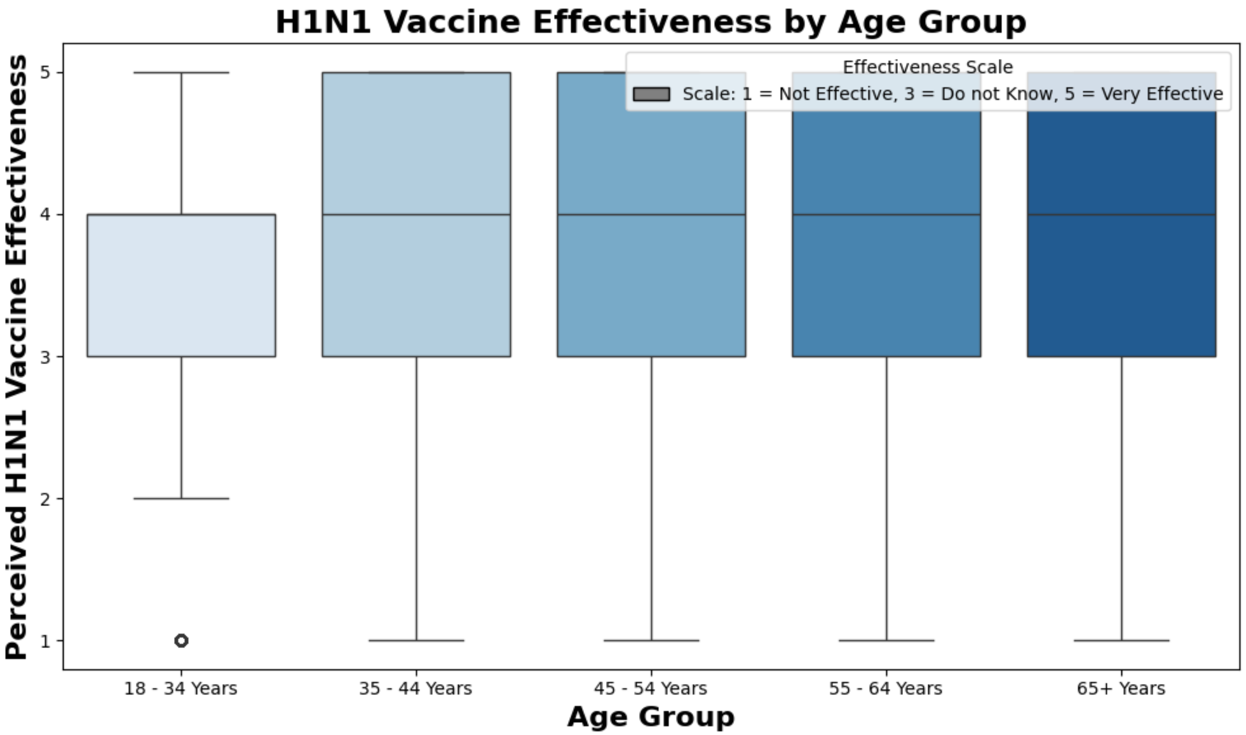
Further Reach

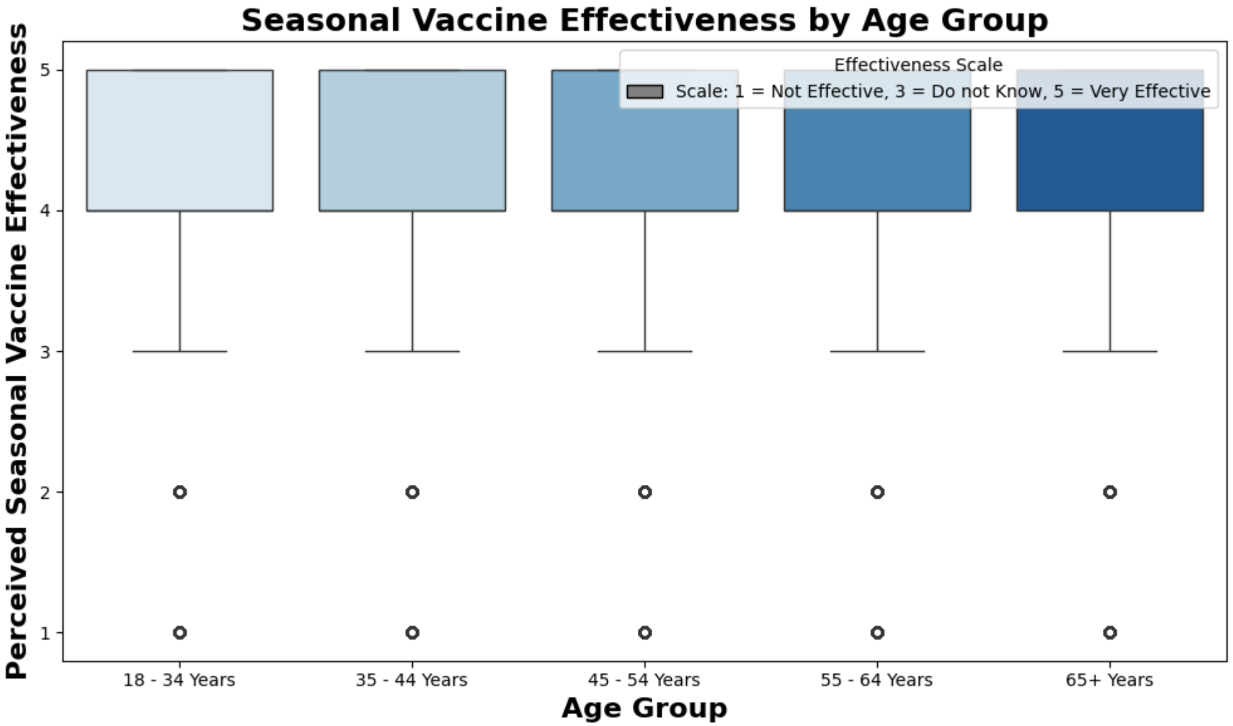
Age Group	Cost Reduction
Strong Provider Recommendation	$ 225 M
School-Based Programs	$ 215 M

All Programs

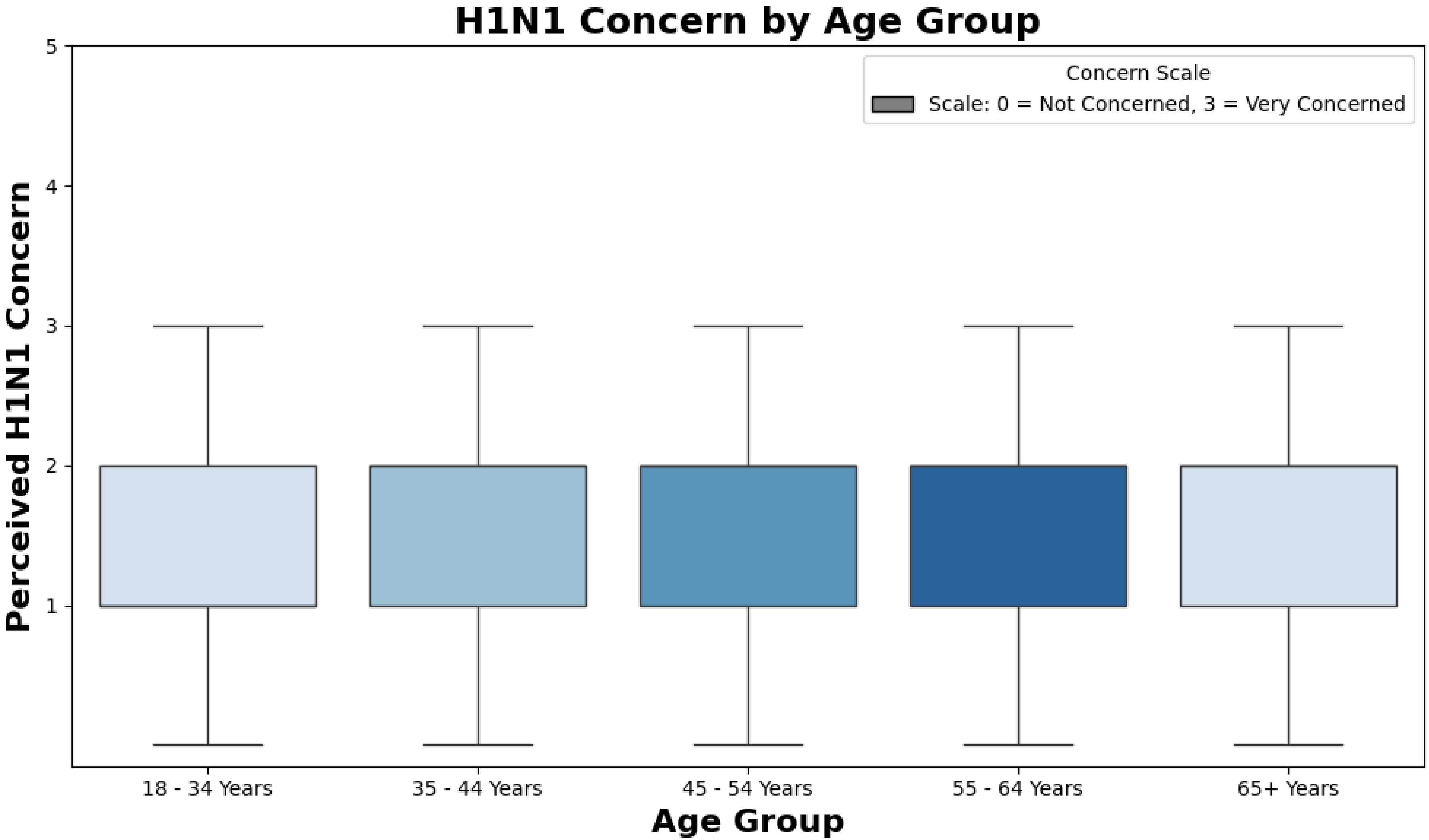
Age Group	Cost Reduction
One-on-One Counseling	$ 210 M
Reminder/Recall Systems	$ 208 M
Strong Provider Recommendation	$ 206 M
School-Based Programs	$ 196 M
Financial Incentives	$ 193 M
Multicomponent Education	$ 217 M

H1N1 & Seasonal Effectiveness By Age Group

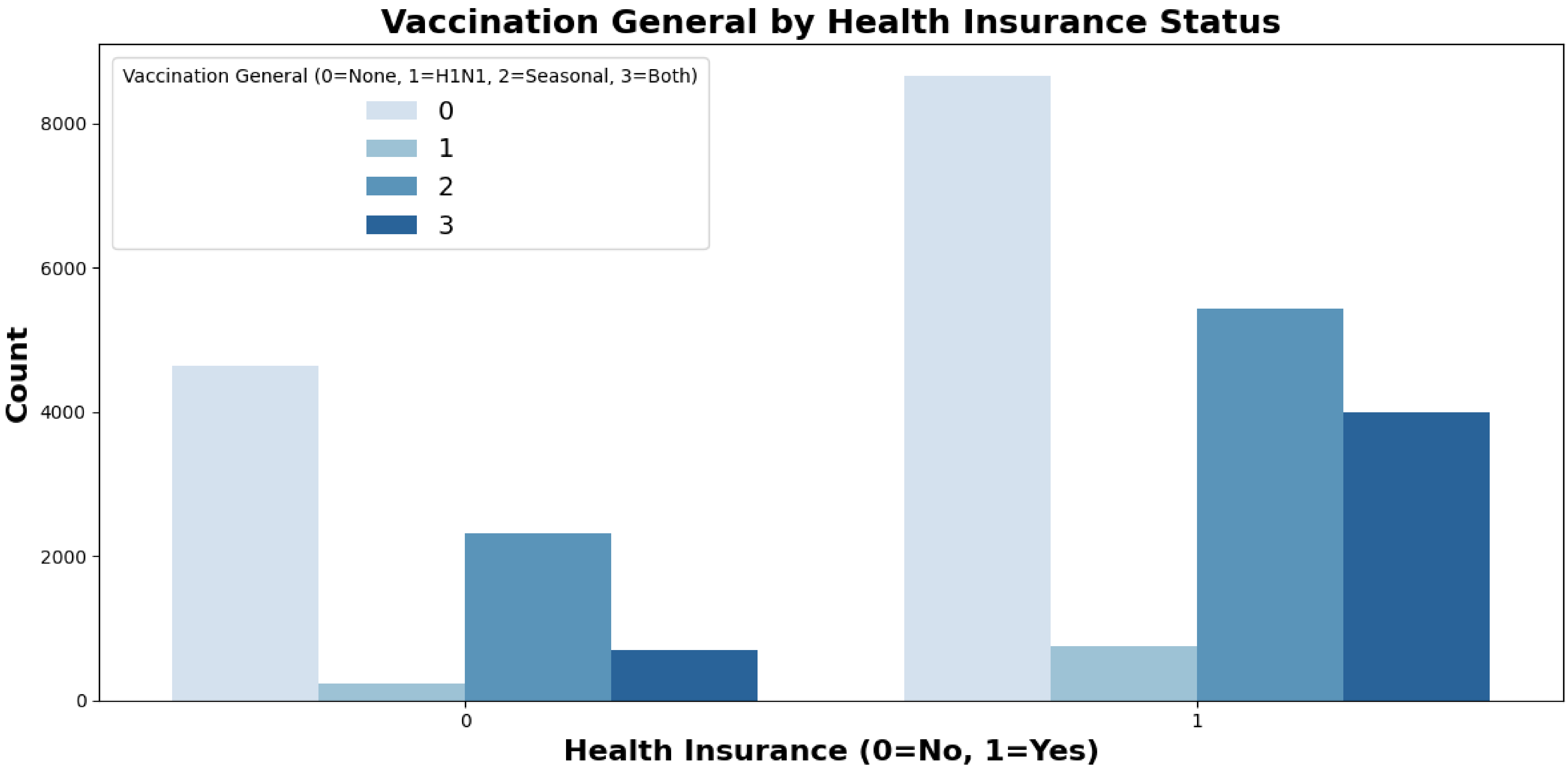




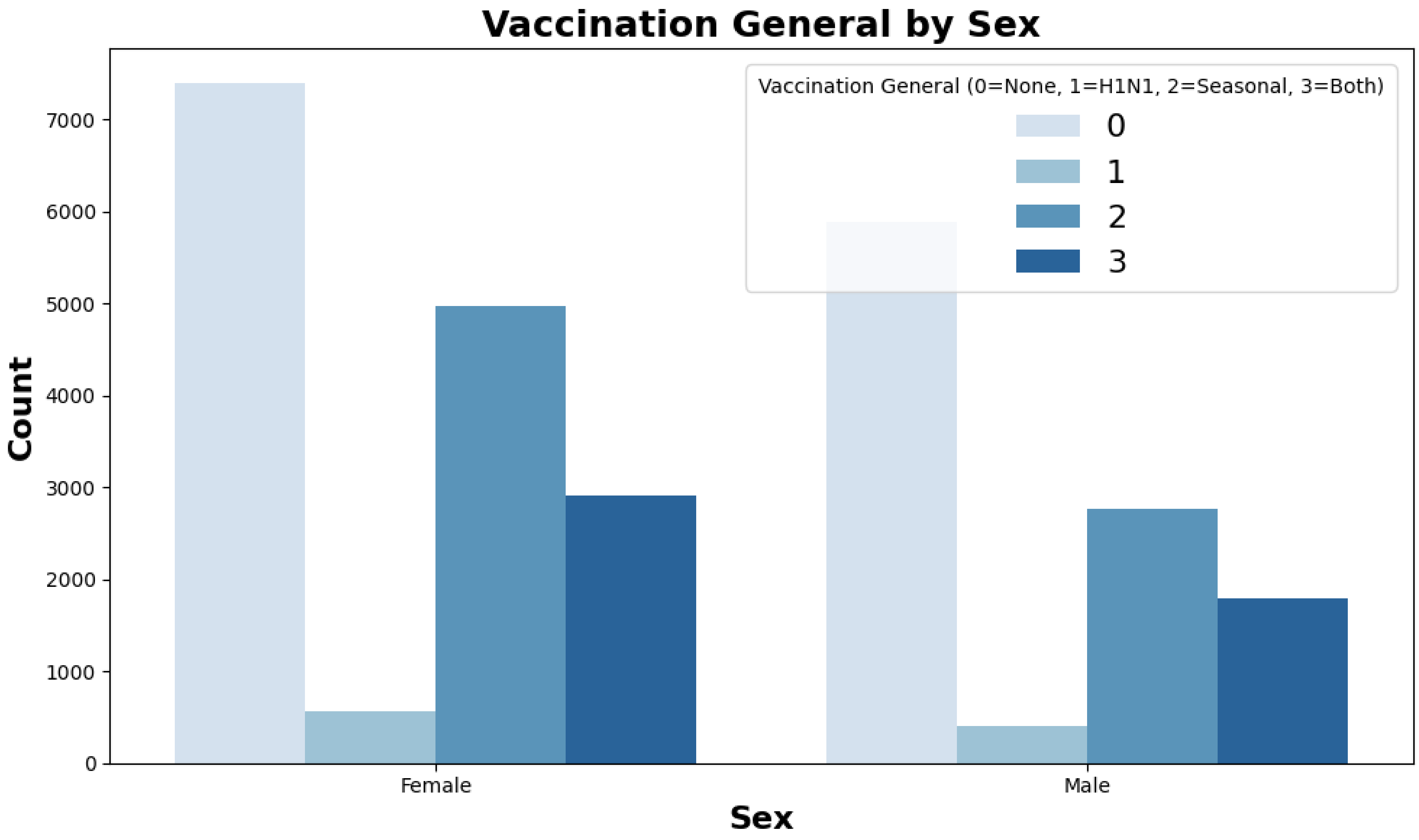
Risk Perception of H1N1 by Age Group



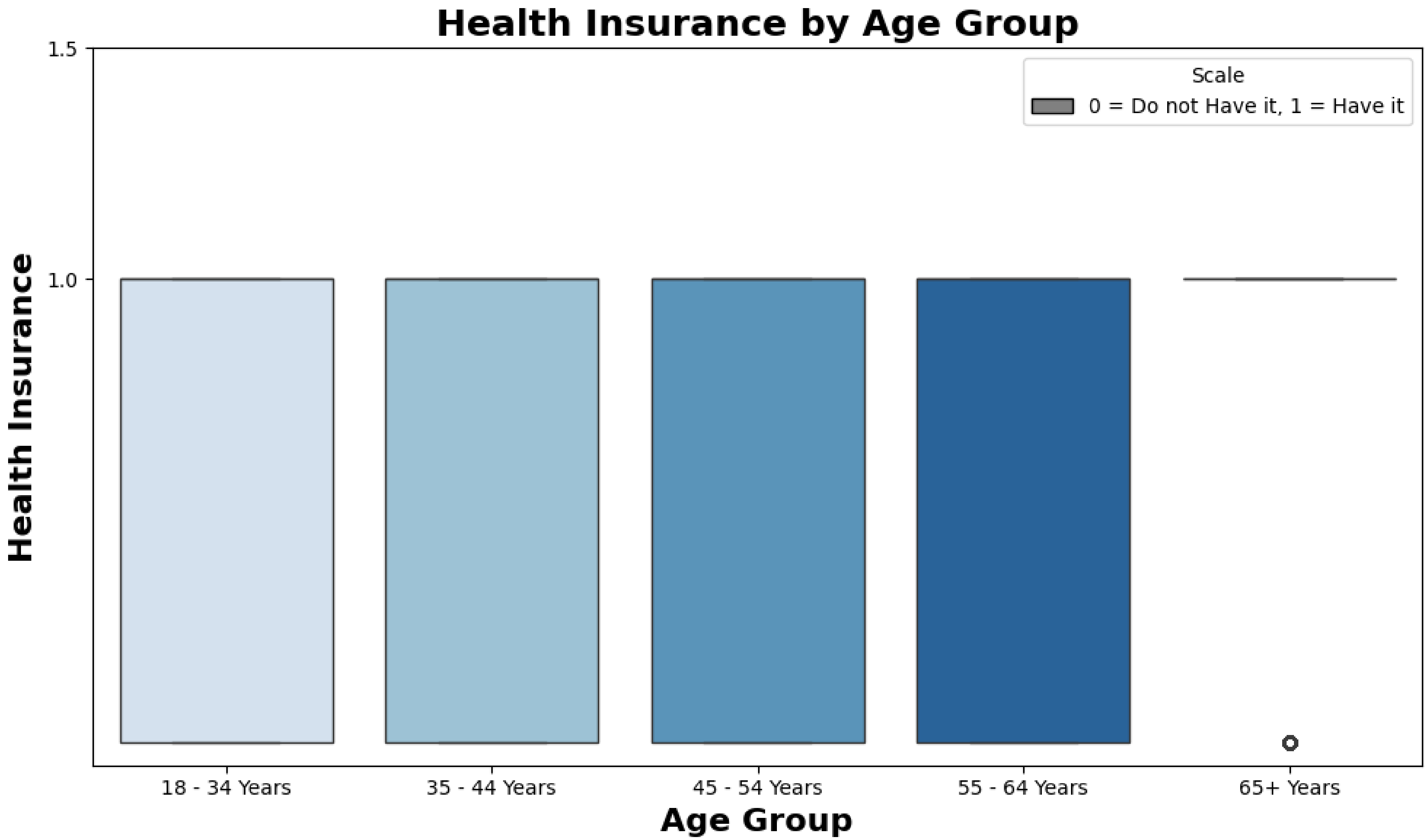
Vaccination Status by Health Insurance Status



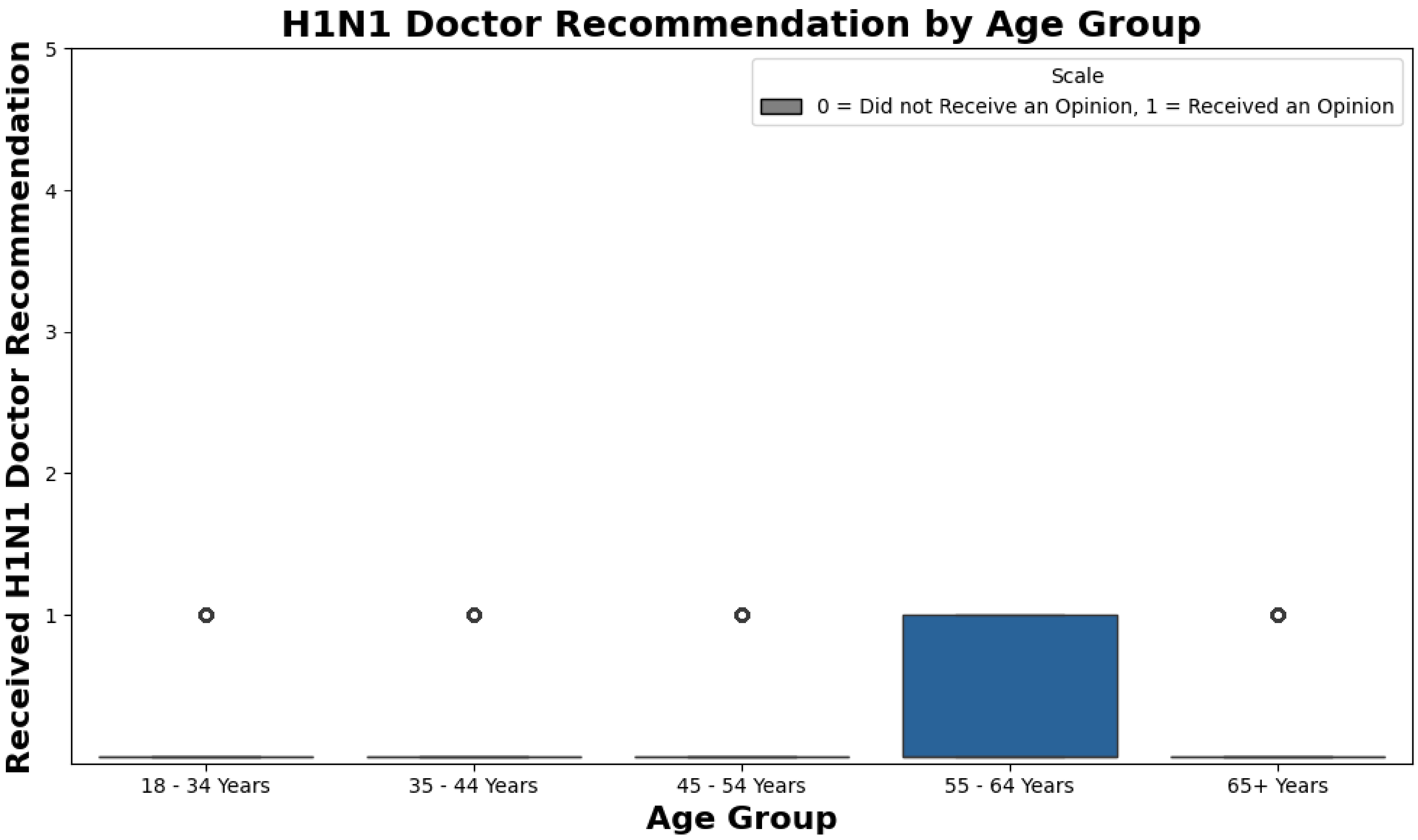
Vaccination Status by Sex



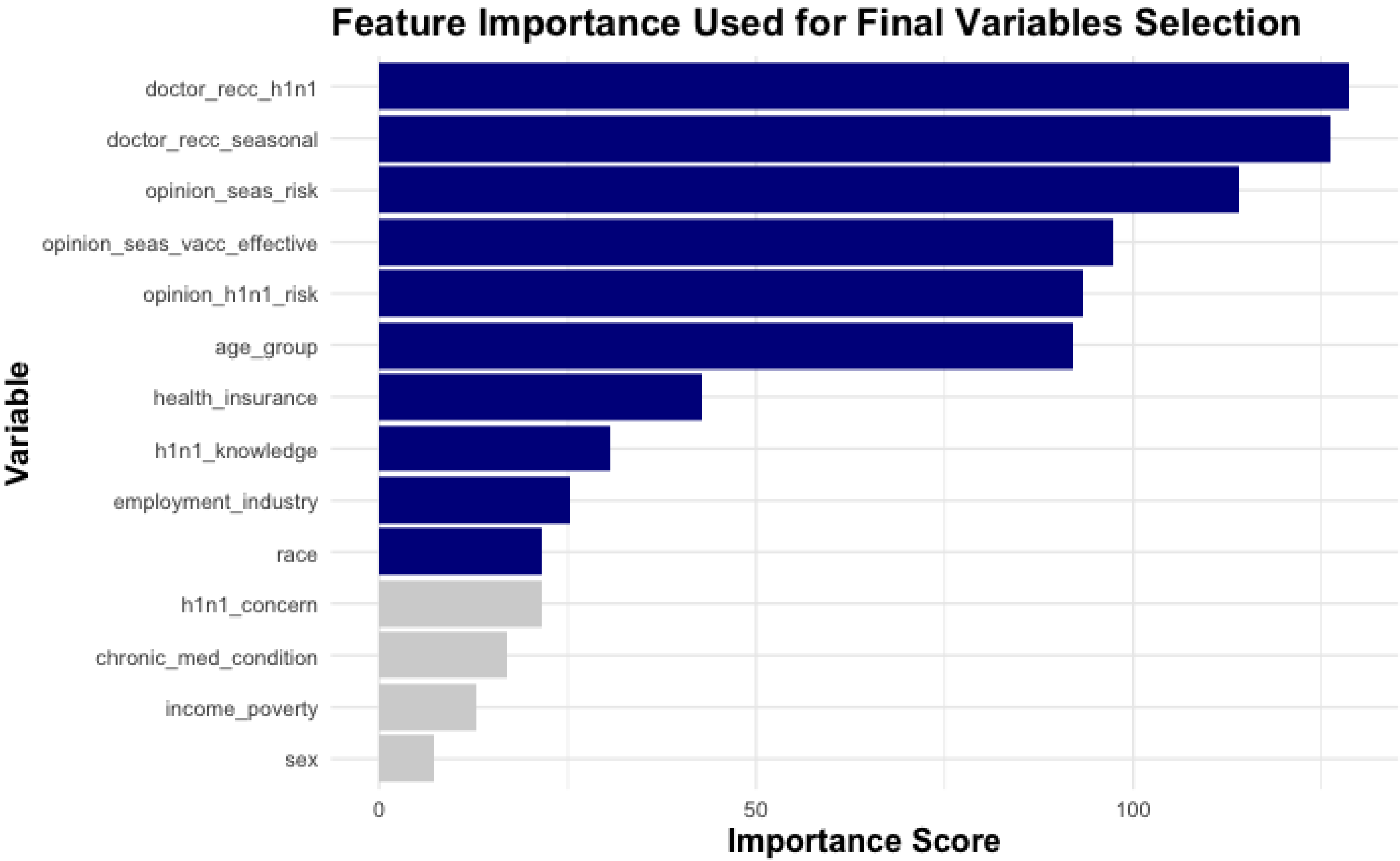
Health Insurance By Age Group



H1N1 Doctor Recommendation By Age Group



Random Forest Feature Importance Used for Final Variable Selection


Feature Importance Used for Final Variables Selection
Variable
doctor_recc_h1n1
doctor_recc_seasonal
opinion_seas_risk
opinion_seas_vacc_effective
opinion_h1n1_risk
age_group
health_insurance
h1n1_knowledge
employment_industry
race
h1n1_concern
chronic_med_condition
income_poverty
sex
Importance Score