J Korean Med Sci. 2021 Aug 9;36(31):e198
https://doi.org/10.3346/jkms.2021.36.e198
eISSN 1598-6357·pISSN 1011-8934

JKMS

Original Article
Medical Informatics

Machine Learning Approach for Active Vaccine Safety Monitoring

Yujeong Kim[1*] **Jong-Hwan Jang**[1*] **Namgi Park**[2] **Na-Young Jeong**[3] **Eunsun Lim**[3] **Soyun Kim**[2] **Nam-Kyong Choi**[3] and **Dukyong Yoon**[1,4]

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Yongin, Korea
[2]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea
[3]Department of Health Convergence, Ewha Womans University, Seoul, Korea
[4]Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin, Korea

Received: Apr 1, 2021
Accepted: Jun 28, 2021

Address for Correspondence:
Dukyong Yoon, MD, PhD
Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 363 Dongbaekjukjeon-daero, Giheung-gu, Yongin 16995, Republic of Korea.
E-mail: dukyong.yoon@yonsei.ac.kr

*Yujeong Kim and Jong-Hwan Jang contributed equally to this study.

ORCID iDs
Yujeong Kim
https://orcid.org/0000-0003-3674-993X
Jong-Hwan Jang
https://orcid.org/0000-0003-3392-822X
Namgi Park
https://orcid.org/0000-0002-2161-5093
Na-Young Jeong
https://orcid.org/0000-0003-2130-1286
Eunsun Lim
https://orcid.org/0000-0003-2130-5682
Soyun Kim
https://orcid.org/0000-0001-5545-2387

ABSTRACT

Background: Vaccine safety surveillance is important because it is related to vaccine hesitancy, which affects vaccination rate. To increase confidence in vaccination, the active monitoring of vaccine adverse events is important. For effective active surveillance, we developed and verified a machine learning-based active surveillance system using national claim data.
Methods: We used two databases, one from the Korea Disease Control and Prevention Agency, which contains flu vaccination records for the elderly, and another from the National Health Insurance Service, which contains the claim data of vaccinated people. We developed a case-crossover design based machine learning model to predict the health outcome of interest events (anaphylaxis and agranulocytosis) using a random forest. Feature importance values were evaluated to determine candidate associations with each outcome. We investigated the relationship of the features to each event via a literature review, comparison with the Side Effect Resource, and using the Local Interpretable Model-agnostic Explanation method.
Results: The trained model predicted each health outcome of interest with a high accuracy (approximately 70%). We found literature supporting our results, and most of the important drug-related features were listed in the Side Effect Resource database as inducing the health outcome of interest. For anaphylaxis, flu vaccination ranked high in our feature importance analysis and had a positive association in Local Interpretable Model-Agnostic Explanation analysis. Although the feature importance of vaccination was lower for agranulocytosis, it also had a positive relationship in the Local Interpretable Model-Agnostic Explanation analysis.
Conclusion: We developed a machine learning-based active surveillance system for detecting possible factors that can induce adverse events using health claim and vaccination databases. The results of the study demonstrated a potentially useful application of two linked national health record databases. Our model can contribute to the establishment of a system for conducting active surveillance on vaccination.

Keywords: Vaccines; Adverse Effects; Postmarketing Product Surveillance; Machine Learning; Cross-over Studies

Nam-Kyong Choi
https://orcid.org/0000-0003-1153-9928
Dukyong Yoon
https://orcid.org/0000-0003-1635-8376

Funding

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT; Ministry of Trade, Industry and Energy; Ministry of Health & Welfare; and Ministry of Food and Drug Safety) (Project Number: 1711138152, KMDF_PR_20200901_0095). This work was also supported by a Government-wide R&D Fund project for infectious disease research (GFID), Republic of Korea (grant number: HG18C0067).

Disclosure

DY is the founder of BUD.on Inc. BUD.on Inc. did not have any role in the study design, analysis, decision to publish, or the preparation of the manuscript. There are no patents, products in development, or marketed products to declare. The other authors declare that they have no competing interests.

Author Contributions

Conceptualization: Yoon D. Data curation and analysis: Kim Y, Jang JH, Park N, Jeong NY. Validation: Kim Y, Jang JH, Park N. Writing - original draft: Kim Y, Jang JH, Park N. Writing - review & editing: Kim Y, Jang JH, Jeong NY, Lim E, Kim S, Choi NK, Yoon D.

INTRODUCTION

As coronavirus disease 2019 (COVID-19) increases in prevalence, the importance of vaccination has increased. Vaccination is essential to achieve herd immunity, and to achieve herd immunity, it is known that at least 70% of the population must be vaccinated.[1,2] However, after vaccination, vaccine adverse events can follow, and these events can cause "vaccine hesitancy," which is directly related to the vaccination rate.[3] If inappropriate information about the adverse events of the vaccine trigger fear of the vaccine, the progress toward herd immunity can be hindered.[4] To reduce vaccine hesitancy, we need to detect the possibility of an adverse effect (which is called a signal) as soon as possible by post-marketing surveillance and provide objective and reliable results to the public after prompt investigation of the detected signals.[5]

Post-marketing surveillance consists of two types: passive and active. Passive surveillance is based on spontaneous reporting systems,[6] which are reporting systems from doctors or patients. Currently, most surveillance relies on passive surveillance,[7] but this approach has several limitations such as under-reporting and reporting bias. South Korea started using passive surveillance systems in 1994, and the Communicable Disease Control Act has required healthcare professionals to report vaccine adverse events since 2001.[8] In contrast, active surveillance actively searches for the adverse events of drugs or vaccines by monitoring existing data such as electronic health records[9,10] or administrative claims data.[11,12] Most importantly, an active surveillance system can detect adverse events more rapidly than a passive system and it helps inform the public about the safety of vaccines.[9] In South Korea, however, an active surveillance system for capturing adverse event information has not yet been established.[7] In a pandemic situation like the current one, catching adverse events early after a large-scale vaccination is important, and a well-developed active surveillance system can form the basis of vaccine safety monitoring.[13]

Awareness of the importance of active surveillance has to the development of several algorithms on large databases,[14] but these approaches still have limitations. Azadeh and Gonzalez conducted an adverse event monitoring study by analyzing the content of social media data mentioning adverse events using the association rule.[15] Botsis et al.[16] used adverse event reports from the Vaccine Adverse Event Reporting System to develop a text mining model that distinguishes between positive and negative reports of anaphylaxis after the H1N1 vaccination. However, adverse event studies based on unstructured text analysis have the disadvantages of being time consuming and labor intensive. To investigate each adverse event, the text must be annotated, and the supervision of a domain expert is essential.[17] Therefore, this approach is not suitable in situations such as COVID-19 vaccination, which requires prompt screening of adverse event signals while overwhelming amounts of new data are generated daily.

Disproportionality analysis, one of most popularly used pre-existing data mining approaches, can be easily applied to big databases retrospectively because of its simplicity,[18] but it may not consider covariates.[19] The TreeScan algorithm can make adjustments to covariates like sex, age, and health plan, but it requires a predefined hierarchical tree structure for the adverse events; the International Classification of Diseases, Ninth Revision, Clinical Modification coding system was used in the previous study.[20] The process that cuts the branches of the tree structure where the ratio of observed-to-expected adverse events is higher seems similar to the process of selecting important features in a decision tree model, and it inspires us to consider the possibility of using the feature importance calculated from machine learning models for screening adverse event signals.

The purpose of this study is to suggest a machine learning-based approach using feature importance for monitoring vaccine adverse events. We established our active surveillance model using elderly flu vaccination records provided by the Korea Disease Control and Prevention Agency (KDCA) and their claim data combined with data from the National Health Insurance Service (NHIS). We demonstrate the reliability of our approach by comparing the results of our model when applied on national claim data with the results of other adverse event databases and the literature for two adverse events: anaphylaxis and agranulocytosis. Moreover, to evaluate its applicability for vaccines, we evaluated the risk of flu vaccination for the elderly on the two health outcomes of interest (HOIs). Our model can be adapted to pandemic situations such as COVID-19 to quickly identify the adverse events of newly developed vaccines.

METHODS

Data description

We used a combined database that consists of a flu vaccination record database for the elderly from the KDCA and a claims database from the NHIS. First, the KDCA extracted the vaccination data and sent them to the NHIS for merging with the claim data. Because these two sets of data contained the resident registration number as the common primary key, the NHIS joined the two datasets using that key and then anonymized the identifier to avoid reidentification. Afterwards, we received the claims data joined with the vaccination records from 2015 to 2019 of this population from the NHIS with anonymized identifiers. Flu vaccination database that we used for our research consists of vaccination data for seniors who are over 65 years of age and received the flu vaccine from 2015 to 2018. The KDCA vaccination data includes the demographic information of the vaccinated and general information related to the vaccination such as the vaccine code, the vaccination time, and lot number of the vaccine. The claims data contains demographic information such as birth year and gender as well as medical treatment data such as diagnoses and prescriptions.

Data preprocessing

We selected anaphylaxis and agranulocytosis as the two HOIs. The Korean Standard Classification of Diseases (KCD-7) codes to determine each HOI are T78.0, T78.2, T80.5, and T88.6 for anaphylaxis[21] and D70 for agranulocytosis.[22]

Data preprocessing consisted of 5 steps. We first extracted the “person identifier,” “prescription identifier,” and “diagnosis date” data for each HOI (except for ruled-out diagnoses) from the NHIS cohort. Second, from the prescription table, we obtained all prescription information for the extracted subjects: “prescription identifier,” “drug code,” and “prescription date.” For the “drug code,” the first 4 digits of the Anatomical Therapeutic Chemical (ATC) code was used. Third, we extracted personal information from the demographic table, which contains information such as the “person identifier,” “birth year,” “living area,” and “income level.” Fourth, we excluded subjects who did not have any demographic information. Finally, we extracted the “person identifier” and “vaccinated date” data for those individuals who were diagnosed with each HOI from the KDCA vaccination database.

Study design

To extract features for use as input to the machine learning model, we adopted case-crossover, which is a major epidemiological design used in vaccine research. If a patient had one HOI diagnosis during the study period, the diagnosis date was used as the index date. If a patient had multiple HOI diagnoses, we treated two consecutive diagnoses as one when the difference in the date of diagnosis was less than 31 days, and used the first diagnosis date as the index date. When the diagnoses were recurrent and more than 31 days apart, we considered recurrent diagnoses as independent HOIs. The minimum interval between diagnoses was set at least six months.

The period of 14 days before the index date was used as the risk window, and we randomly chose 14 days as control windows (excluding the days in the risk windows and washout periods). Next, we extracted all the prescription and vaccination records corresponding with each window. Windows that did not include any records in each period, we excluded from the research. If there are no prescription or vaccination records in the risk window, only the control windows were used. For anaphylaxis, 7,332 people out of 14,047 had control windows only, and for agranulocytosis, this was true for 5,712 people out of 28,005. The overall research process is summarized in **Figs. 1** and **2**.

Machine learning model construction

We developed a random forest-based model for this study. The extracted prescription and vaccination records were used as the input values, and the targeted outcome was a HOI (i.e., whether the extracted information was from the risk or control window). The input dataset was randomly split into training (80%) and test (20%) sets. To avoid overfitting and obtain generalized performance results, we performed bootstrapping 100 times. Evaluations were performed on both the training and test sets. Sensitivity, specificity, the f1-score, accuracy, and the area under the curve (AUROC) were calculated as performance parameters.

Feature importance was evaluated to determine features were important for detecting each HOI. We obtained the overall feature importance values from each bootstrapping result. We defined the feature importance ratio as the ratio of each feature's average importance value to the average of all feature importance (except those with zero importance) over 100

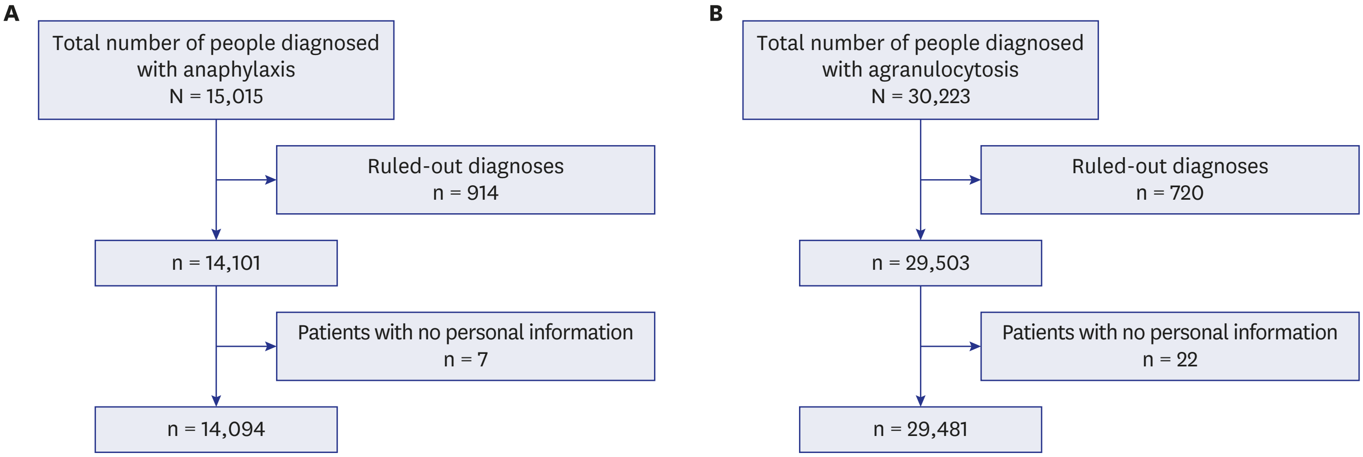
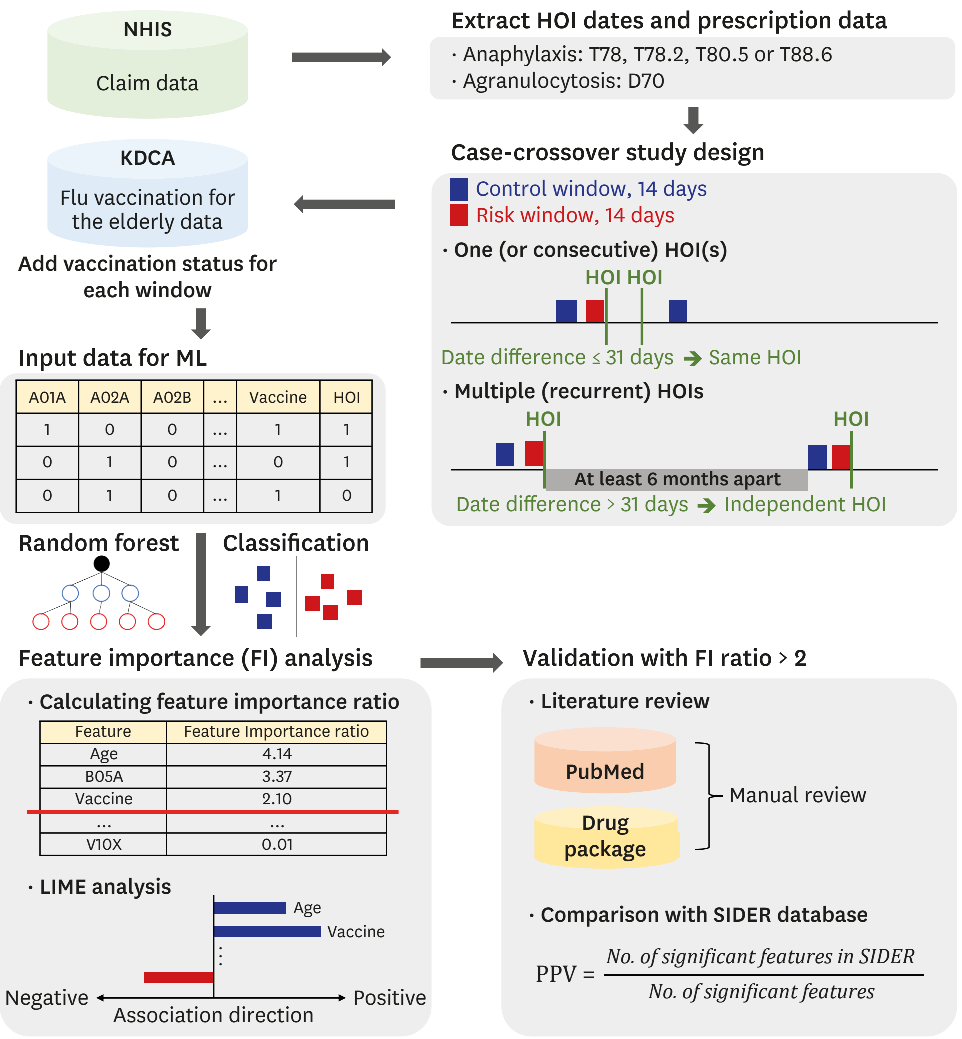

A
Total number of people diagnosed with anaphylaxis N = 15,015
Ruled-out diagnoses n = 914
n = 14,101
Patients with no personal information n = 7
n = 14,094
B
Total number of people diagnosed with agranulocytosis N = 30,223
Ruled-out diagnoses n = 720
n = 29,503
Patients with no personal information n = 22
n = 29,481

Fig. 1. Flow chart of the study. (**A**) Anaphylaxis. (**B**) Agranulocytosis. Among the whole population, the numbers of people who were diagnosed with anaphylaxis or agranulocytosis at least once were 15,015 and 30,223, respectively. We selected people who did not have ruled-out diagnoses and for whom demographic information was available. After adapting exclusion criteria, the final sample size was 14,094 for anaphylaxis and 29,481 for agranulocytosis.


NHIS
Claim data
Extract HOI dates and prescription data
· Anaphylaxis: T78, T78.2, T80.5 or T88.6
· Agranulocytosis: D70
Case-crossover study design
🟦 Control window, 14 days
🟥 Risk window, 14 days
· One (or consecutive) HOI(s)
HOI HOI
Date difference ≤ 31 days → Same HOI
· Multiple (recurrent) HOIs
HOI
HOI
At least 6 months apart
Date difference > 31 days → Independent HOI
KDCA
Flu vaccination for the elderly data
Add vaccination status for each window
Input data for ML
Random forest
Classification
Feature importance (FI) analysis
· Calculating feature importance ratio
· LIME analysis
Age
Vaccine
Negative
Positive
Association direction
Validation with FI ratio > 2
· Literature review
PubMed
Drug package
Manual review
· Comparison with SIDER database

A01A	A02A	A02B	...	Vaccine	HOI
1	0	0	...	1	1
0	1	0	...	0	1
0	1	0	...	1	0

Feature	Feature Importance ratio
Age	4.14
B05A	3.37
Vaccine	2.10
...	...
V10X	0.01

$$\text{PPV} = \frac{\textit{No. of significant features in SIDER}}{\textit{No. of significant features}}$$

Fig. 2. Research workflow for developing an active surveillance system. First, total HOI diagnosis dates were extracted using the NHIS claim dataset. The period of 14 days before the HOI diagnosis was set as a risk window. The control window was randomly selected for 14 days excluding the risk window and washout period. If a HOI occurred several times, the HOI that re-occurred within 31 days was considered as the same and a continuous event of the previous HOI. In order to ensure independence between HOIs, a risk window was defined only for recurrent HOI events where the interval between HOIs was more than 6 months apart. Second, the prescription and vaccination information that occurred in each window was collected. If there was no prescription and vaccination information in the window, the window was removed. At the final step, the HOI prediction model was learned using the prescription and vaccination information in each window. After that, using the feature importance ratio and LIME analysis of the model, a suspected drug or vaccine that could cause the HOI was determined.
FI = feature importance, HOI = health outcome of interest, KDCA = Korea Disease Control and Prevention Agency, LIME = Local interpretable Model-agnostic Explanation, NHIS = National Health Insurance Service, SIDER = Side Effect Resource database.

bootstrapping trials. We treat features with a feature importance ratio of more than 2 as candidates that indicate adverse events.

$$\text{Feature importance ratio} = \frac{\text{Average of each feature's importance}}{\text{Average of whole features' importances}}$$

Validation of the detected signals

To validate our study results, we reviewed the literature manually and compared the study results with the Side Effect Resource (SIDER) database. In the literature review, we searched for an association between the HOI and important features from our study results and calculated the ratio of the number of important features with references to the total number of important features. Using the SIDER database, we calculated the positive predictive value (PPV), which is the ratio of the number of important features that had a record in SIDER to the total number of important features.

Local interpretable Model-agnostic Explanation (LIME) analysis

LIME is a method to check which variables influenced the prediction results and how much they did so in a nonlinear machine learning prediction model, which is difficult to interpret.[23] The random forest model provides the feature importance, which indicates the influence of a feature on the prediction result, but it does not indicate whether the variable has a positive or negative influence on the prediction. For example, if the feature importance of vaccination is 0.9 in a random forest model, vaccination may be informative in distinguishing people without adverse events, but there is no information whether this effect is positive or negative on the HOI. The results of LIME could provide the direction of association between the features and HOI by estimating the change in prediction results caused by changing the local values of each feature.

Software

In our study, the data preprocessing and model development scripts were written in Python version 3.6 using the scikit-learn Python package. We tuned the hyperparameters of our model using scikit-learn's RandomForestClassifier and RandomizedSearchCV functions.[24] To interpret the machine learning results, the LIME package was used.[23]

Ethics statement

This study was approved by the Institutional Review Board (IRB) of Yonsei University Severance Hospital (IRB No. 9-2021-0019). No informed consent was required from patients due to the nature of public data from NHIS and KCDA.

RESULTS

Table 1 summarizes the patients' demographic information for the total dataset and each HOI cohort at the first diagnosis. Because we analyzed flu-vaccinated elderly subjects, the mean age of each anaphylaxis and agranulocytosis cohort was 73 and 71 years, respectively, and the numbers of each HOI occurring at least once were 8,335 and 20,673, respectively. In addition, the numbers of data with vaccination in each HOI dataset were 35,536 and 17,216, respectively.

Performance

Table 2 presents the performance of the surveillance model for each HOI. On the training data, the performance metrics were all over 0.90, which means the model was trained well to predict whether each HOI would occur or not. For the test data for anaphylaxis, it scored about 0.69 in the AUROC, accuracy, recall, and precision metrics and 0.67 in f1-score on the

Table 1. Baseline characteristics at first HOI diagnosis

Name	Total dataset[a]	Anaphylaxis	Agranulocytosis
No. of patients (%)[b]	5,544,150	14,046 (0.25)	28,005 (0.51)
Age (Mean ± SD)	73 ± 7	71 ± 6	73 ± 6
Sex			
No. of Male	2,397,324	7,236	13,309
No. of Female	3,146,826	6,810	14,696

HOI = health outcome of interest, SD = standard deviation.
[a]Claim dataset joined with flu vaccination records. [b]Percentage of patients in the total dataset who had ever experienced a HOI (anaphylaxis or agranulocytosis).

Table 2. Performance of the machine learning model for each health outcome of interest

Dataset	Metric	Anaphylaxis performance	Agranulocytosis performance
Training set	Recall	0.940	0.949
	Precision	0.945	0.953
	f1-score	0.939	0.949
	Accuracy	0.940	0.949
	AUROC	0.995	0.994
Test set	Recall	0.698	0.729
	Precision	0.695	0.727
	f1-score	0.670	0.726
	Accuracy	0.698	0.729
	AUROC	0.697	0.770

AUROC = area under the receiver operating curve.

test set. For agranulocytosis, the recall, precision, f1-score, and accuracy metrics of the model using the test set scored around 0.72.

Feature importance list

The importance of the features of the model for predicting HOIs and their feature importance ratios are presented in **Table 3**. Whole features with an importance ratio over 2 are listed in the Supplementary Materials. Twenty-seven features were found to be important because their feature importance ratio was over 2. Age had the highest feature importance ratio, and sex was the 6th most important feature. The status of flu vaccination was also significant (importance ratio: 3.53) and ranked 11th. Among drug-related features, drugs for peptic ulcer and for gastro-esophageal reflux disease (GORD), nonsteroidal anti-inflammatory drugs (NSAIDs), antihistamines, and corticosteroids ranked high. In the results of agranulocytosis, 25 features had a feature importance ratio of over 2. The most important feature was age, with a feature importance ratio of 18.4, and the status of vaccination was ranked 38th with a feature importance ratio less than 2 (1.49). Blood substitutes, solutions, antihistamines, drugs for GORD, and corticosteroids ranked high as drugs (in that order).

Table 3. Feature importance ratios for the top-10 features and vaccine

HOI	Feature	Feature importance ratio	Definition
Anaphylaxis	Age	30.31	AGE
	A02B	6.97	DRUGS FOR PEPTIC ULCER AND GASTRO-OESOPHAGEAL REFLUX DISEASE (GORD)
	M01A	5.82	ANTI-INFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS
	R06A	5.64	ANTIHISTAMINES FOR SYSTEMIC USE
	H02A	5.52	CORTICOSTEROIDS FOR SYSTEMIC USE, PLAIN
	Sex	5.31	SEX
	N02A	4.43	OPIOIDS
	A03F	4.33	PROPULSIVES
	N01B	3.78	ANESTHETICS, LOCAL
	M09A	3.75	OTHER DRUGS FOR DISORDERS OF THE MUSCULO-SKELETAL SYSTEM
	Vaccine	3.53	VACCINATION STATUS
Agranulocytosis	Age	18.43	AGE
	B05X	14.33	I.V. SOLUTION ADDITIVES
	B05B	10.82	I.V SOLUTIONS
	A04A	6.17	ANTIEMETICS AND ANTINAUSEANTS
	A02B	5.84	DRUGS FOR PEPTIC ULCER AND GASTRO-OESOPHAGEAL REFLUX DISEASE (GORD)
	H02A	5.64	CORTICOSTEROIDS FOR SYSTEMIC USE, PLAIN
	N02A	4.51	OPIOIDS
	Sex	4.04	SEX
	A03F	4.02	PROPULSIVES
	L01X	3.64	OTHER ANTINEOPLASTIC AGENTS
	...	...	...
	Vaccine	1.49	VACCINATION STATUS

Validation of the important features

Literature evidence

In anaphylaxis, among the 27 features that had an importance ratio of over 2, 19 features (70%) were mentioned as having an association with anaphylaxis in the literature such as scientific journals (see **Supplementary Table 1** for details). For example, beta-lactam antibiotics, NSAIDs, opioids, and neuromuscular blocking agents are considered the most common anaphylaxis-inducing drugs,[25] and they were included in our study results. Montañez et al.[26] noted that the presence of concomitant diseases (asthma, mastocytosis, and cardiovascular diseases) could increase the risk of anaphylaxis, and the drugs used for cardiovascular diseases (blood glucose lowering drugs, lipid modifying agents, and angiotensin-II antagonists) or allergy (antihistamines and corticosteroids) were also included in the study results.

Among 25 features that have an importance ratio of over 2 in agranulocytosis, 19 features (76%) were already mentioned (or remarked on) in the literature such as scientific journals, public documents, or drug packages (see **Supplementary Table 2** for details). For example, Schweizer et al.[27] reported that 23.8% of patients who had a high risk of prostate cancer had febrile neutropenia after taking leuprolide (ATC code: L02AE02). A case report mentioned a patient who took tamsulosin (ATC code: G04CA02) at 0.4 mg/day with other drugs including octreotide and was diagnosed with neutropenia.[28]

Comparison with SIDER data

Among 48 drug-related features which had a feature importance ratio of over 1, 38 features including NSAIDs, anti-infectives, and antidepressants were also listed in SIDER as drugs that can induce anaphylaxis (PPV 79.17%). For the 24 drug-related features that had an importance of over 2, 17 of them were listed in SIDER (PPV 70.83%) (**Fig. 3A**).

Among the 54 drug-related features that had an importance ratio of over 1, 36 features including antinauseants, antihistamines and antidepressants were also listed in SIDER as drugs that can induce agranulocytosis or neutropenia (PPV 66.67%). Of the 23 features that have an importance of over 2, excluding age and sex, 16 of them were listed in SIDER (PPV 69.57%) (**Fig. 3B**).

LIME analysis results

According to the LIME analysis, 19 of 27 features (70%) had positive relationships with anaphylaxis. Vaccination had a high feature importance ratio of 3.53 and a positive relationship with anaphylaxis (**Fig. 4A**). Moreover, 17 out of 25 features (68%) had positive relationships, which means each feature can help induce agranulocytosis. Vaccination status had a positive relationship with agranulocytosis, but the importance ratio was 1.49 (**Fig. 4B**).

DISCUSSION

Our study demonstrates that the machine learning-based vaccine adverse event monitoring system was able to detect which features affected agranulocytosis and anaphylaxis using the elderly flu vaccine cohort data provided by the NHIS and KCDA. The model was trained well, with a training data AUROC of over 99% and predicted HOI occurrences on the test data with AUROC values of around 70% for both outcomes. We proved that most features with an importance ratio of over 2 were related to the occurrences of each HOI by investigating whether each feature was mentioned in the scientific journals or with respect to drug packages. For features with an importance ratio over 2, the PPVs were 69.6% and 70.8% for

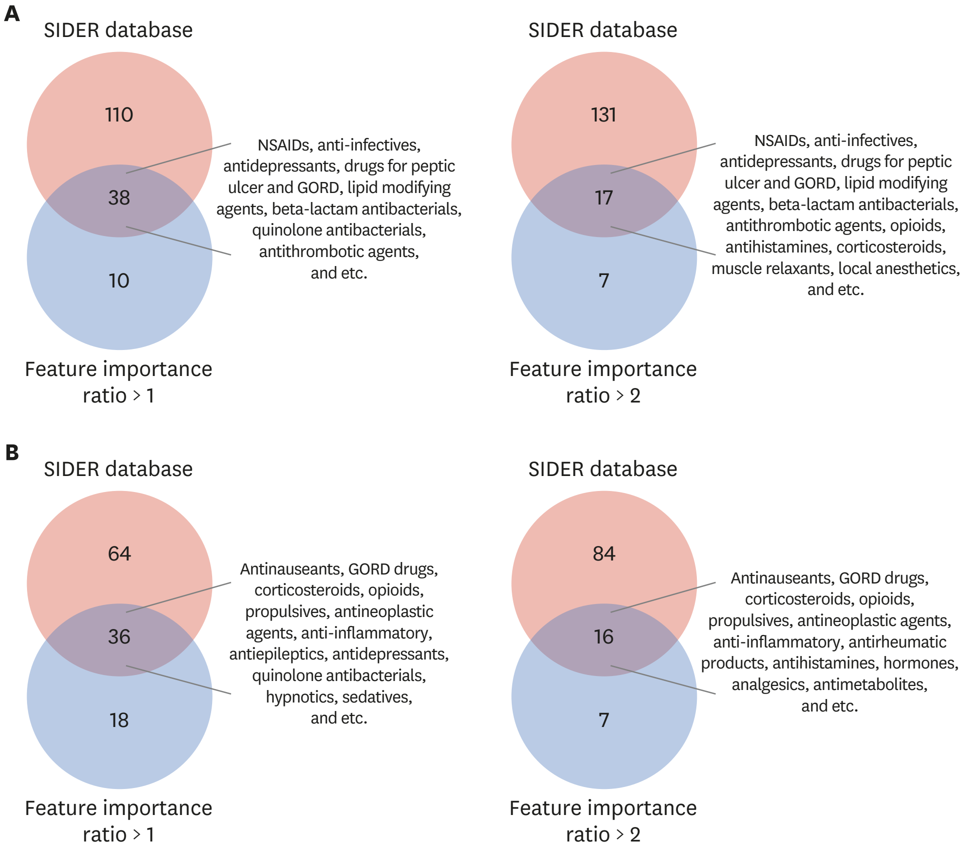

A
SIDER database
110
38
10
Feature importance ratio > 1
NSAIDs, anti-infectives, antidepressants, drugs for peptic ulcer and GORD, lipid modifying agents, beta-lactam antibacterials, quinolone antibacterials, antithrombotic agents, and etc.
SIDER database
131
17
7
Feature importance ratio > 2
NSAIDs, anti-infectives, antidepressants, drugs for peptic ulcer and GORD, lipid modifying agents, beta-lactam antibacterials, antithrombotic agents, opioids, antihistamines, corticosteroids, muscle relaxants, local anesthetics, and etc.
B
SIDER database
64
36
18
Feature importance ratio > 1
Antinauseants, GORD drugs, corticosteroids, opioids, propulsives, antineoplastic agents, anti-inflammatory, antiepileptics, antidepressants, quinolone antibacterials, hypnotics, sedatives, and etc.
SIDER database
84
16
7
Feature importance ratio > 2
Antinauseants, GORD drugs, corticosteroids, opioids, propulsives, antineoplastic agents, anti-inflammatory, antirheumatic products, antihistamines, hormones, analgesics, antimetabolites, and etc.

Fig. 3. Comparison of SIDER-listed features and features with high importance ratios. (**A**) Anaphylaxis. (**B**) Agranulocytosis. The Venn diagrams show the number of features in each result, and the intersection includes those features that are not only already listed in the SIDER database but also have an importance ratio of over 1 (left) or 2 (right).
GORD = gastro-esophageal reflux disease, NSAID = nonsteroidal anti-inflammatory drug, SIDER = Side Effect Resource database.

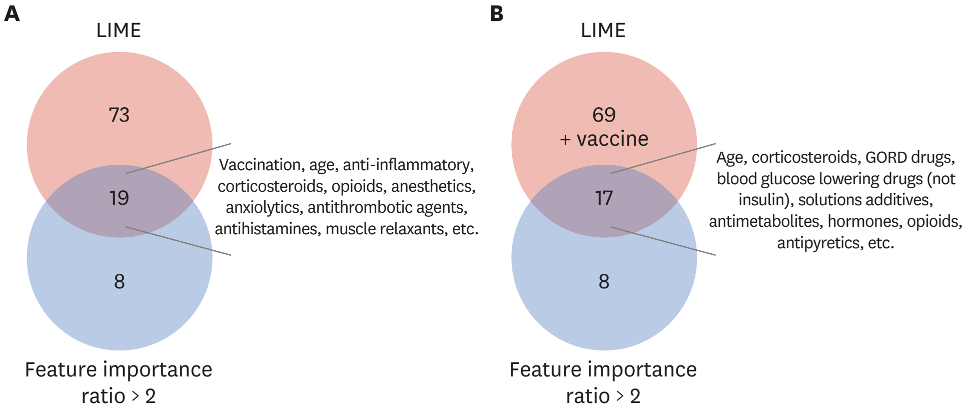

A
LIME
73
19
8
Feature importance ratio > 2
Vaccination, age, anti-inflammatory, corticosteroids, opioids, anesthetics, anxiolytics, antithrombotic agents, antihistamines, muscle relaxants, etc.
B
LIME
69
+ vaccine
17
8
Feature importance ratio > 2
Age, corticosteroids, GORD drugs, blood glucose lowering drugs (not insulin), solutions additives, antimetabolites, hormones, opioids, antipyretics, etc.

Fig. 4. Comparison of LIME positive features and features with an importance ratio of over 2. The Venn diagrams show the number of features in each result and the intersection includes those features that are not only LIME positive but also have an importance ratio of over 2 (**A**) for anaphylaxis and (**B**) for agranulocytosis.
GORD = gastro-esophageal reflux disease, LIME = Local interpretable Model-agnostic Explanation.

agranulocytosis and anaphylaxis, respectively, when compared with data from the SIDER database. Vaccination status was considered to be an important feature in anaphylaxis surveillance results. This result is consistent with the fact that anaphylaxis is known to be an adverse reaction of the flu vaccine.[29] Overall, this indicates that our prediction model, which is a machine learning-based surveillance method that is suitable for big data analysis, has the potential to detect important features that correspond with each HOI.

Our model can be used as a vaccine surveillance system because of its strengths related to surveillance. First, researchers can easily determine lists of suspected factors that can cause a HOI. Traditional statistical methods must calculate all the correlations between each factor and target, but our model can compare whole factors including all kinds of drugs, vaccination status, and demographic features simultaneously for the same HOI. The matching of adverse events and target drugs is often dependent on experts with domain knowledge or passive reporting systems. However, because the drug development process has accelerated and numerous drugs are now on the market, it is difficult for experts to identify HOI-drug pairs. Our research can be effectively used to proactively determine suspicious HOI-drug pairs. Moreover, the random forest model used in this study is known to have better performance and speed than the regression model as the amount of data and the number of features increases, which is more suitable for the analysis of big data.[30]

Our model also demonstrated the potential for determining suspected drug-HOI pairs considering drug–drug interactions. When calculating feature importance, it considers all drugs that were administered within the risk period. For example, ketorolac tromethamine, which is called Toradol, is an NSAID for treating severe pain.[31] The ATC group of ketorolac is M01A, and that of tromethamine is B05B (as a blood substitute). There is no relationship between each drug and agranulocytosis, but the Food and Drug Administration (FDA) has said that the mixture of two drugs can cause agranulocytosis as a side effect.[31] This indicates that our results can imply drug–drug interactions.

Our model can detect suspicious drugs that the SIDER database does not contain. Because the SIDER database contains drug-HOI pairs up to 2015, there may be adverse reactions that had not been discovered at that time. We searched for suspicious drugs, which are called false positives and had a high importance ratio in our model. Therefore, we tried to obtain the reliability of the results through a literature search on false positive analyses and the results. First, our model outputs that A03F (propulsive) had a high importance ratio for inducing anaphylaxis, but this is not listed in the SIDER database. However, we found that Dhakal et al.[32] mentioned anaphylaxis as a very rare side effect of Domperidone. Second, our model said G04C (drugs used in benign prostatic hypertrophy) was important for inducing agranulocytosis, but this is also not listed in the SIDER database. We found there was a case in which tamsulosin induced neutropenia and thrombocytopenia.[33] These false positive related evidences make our model more powerful.

Another strength of our model is that it can screen for all factors that can induce a targeted HOI. In other words, periodic screening is possible by determining a HOI with a high fatality rate and a high risk, and this model can be effectively used by drug-related government departments. We present a new surveillance process accordingly: After listing all suspected drugs or vaccines that could cause a HOI using our model, an expert can select a possible HOI-drug pair from among the candidate pairs output by the model and perform a precise statistical analysis. Finally, if a significant result is found in the statistical analysis, a clinical verification can be performed through an epidemiological investigation. By providing suspicious drug-HOI pairs, surveillance can be performed faster and more effectively than current approaches, which rely on expert knowledge or passive reporting.

There are some limitations of this study that should be declared. First, machine learning-based models are more effective when the number of data and features are large but the data used in this study were not large enough to confirm the advantages of machine learning. In addition, because the data of elderly people who had been vaccinated for flu vaccine was used, the age range of the cohort was limited. However, through this study, we confirmed that the machine learning-based active surveillance monitoring system for adverse events can be effective. Second, feature importance analysis does not guarantee a causal relationship between drugs and adverse events. For example, among the factors included in the feature importance for the results of agranulocytosis, there were cases in which adjuvants such as fluids were present. However, in the case of adjuvants, their purpose and usage are clear, so the researcher can easily filter them out. The purpose of this study is not to investigate the exact causal relationship between drugs and HOI, but to quickly monitor adverse event candidates. Moreover, for screening purposes, the method proposed in this study could be effectively used to determine drug prescription patterns that are related to HOI. Third, our method can provide PPVs for predicting HOI induced by drugs, but it cannot provide sensitivity because we cannot know the complete list of drug-induced adverse reactions. This is a fundamental limitation of surveillance studies that aim to detect unknown adverse events. However, by providing a false positive analysis, we obtain the reliability of the research.

To our knowledge, this is the first study to develop a machine learning-based surveillance system for detecting suspicious factors that can induce adverse events using a nation-wide insurance claim and vaccination database. The results of the study demonstrated that our model can list all the factors related to a HOI simultaneously. We expect that our model will help to make the adverse event surveillance process more efficient.

SUPPLEMENTARY MATERIALS

Supplementary Table 1

List of features that have a feature importance ratio over 2 and each feature's evidence related to anaphylaxis

Click here to view

Supplementary Table 2

List of features that have a feature importance ratio over 2 and each feature's evidence related to agranulocytosis

Click here to view

REFERENCES

1. Rubin R. Difficult to determine herd immunity threshold for COVID-19. *JAMA* 2020;324(8):732.
PUBMED | CROSSREF

2. Jung J. Preparing for the coronavirus disease (COVID-19) vaccination: evidence, plans, and implications. *J Korean Med Sci* 2021;36(7):e59.
PUBMED | CROSSREF

3. Salmon DA, Dudley MZ, Glanz JM, Omer SB. Vaccine hesitancy: causes, consequences, and a call to action. *Vaccine* 2015;33 Suppl 4:D66-71.
PUBMED | CROSSREF

4. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat Hum Behav* 2021;5(3):337-48.
PUBMED | CROSSREF

5. Dubé E, Laberge C, Guay M, Bramadat P, Roy R, Bettinger J. Vaccine hesitancy: an overview. *Hum Vaccin Immunother* 2013;9(8):1763-73.
PUBMED | CROSSREF

6. Takahashi H, Pool V, Tsai TF, Chen RT; The VAERS Working Group. Adverse events after Japanese encephalitis vaccination: review of post-marketing surveillance data from Japan and the United States. *Vaccine* 2000;18(26):2963-9.
PUBMED | CROSSREF

7. Jeong NY, Park S, Lim E, Choi NK. An introduction of the active vaccine safety surveillance system in foreign countries. *J Health Info Stat.* 2019;44(4):317-30.
CROSSREF

8. Choe YJ, Bae GR. Management of vaccine safety in Korea. *Clin Exp Vaccine Res* 2013;2(1):40-5.
PUBMED | CROSSREF

9. Davis RL, Kolczak M, Lewis E, Nordin J, Goodman M, Shay DK, et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology* 2005;16(3):336-41.
PUBMED | CROSSREF

10. Yih WK, Kulldorff M, Fireman BH, Shui IM, Lewis EM, Klein NP, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. *Pediatrics* 2011;127 Suppl 1:S54-64.
PUBMED | CROSSREF

11. Brown JS, Kulldorff M, Chan KA, Davis RL, Graham D, Pettus PT, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiol Drug Saf* 2007;16(12):1275-84.
PUBMED | CROSSREF

12. Huh K, Kim YE, Radnaabaatar M, Lee DH, Kim DW, Shin SA, et al. Estimating baseline incidence of conditions potentially associated with vaccine adverse events: a call for surveillance system using the Korean National Health Insurance Claims Data. *J Korean Med Sci* 2021;36(9):e67.
PUBMED | CROSSREF

13. Lee GM, Romero JR, Bell BP. Postapproval Vaccine Safety Surveillance for COVID-19 Vaccines in the US. *JAMA* 2020;324(19):1937-8.
PUBMED | CROSSREF

14. Huang YL, Moon J, Segal JB. A comparison of active adverse event surveillance systems worldwide. *Drug Saf* 2014;37(8):581-96.
PUBMED | CROSSREF

15. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc* 2015;22(3):671-81.
PUBMED | CROSSREF

16. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *J Am Med Inform Assoc* 2011;18(5):631-8.
PUBMED | CROSSREF

17. Jeon E, Kim Y, Park H, Park RW, Shin H, Park HA. Analysis of adverse drug reactions identified in nursing notes using reinforcement learning. *Healthc Inform Res* 2020;26(2):104-11.
PUBMED | CROSSREF

18. Lindquist M, Ståhl M, Bate A, Edwards IR, Meyboom RH. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. *Drug Saf* 2000;23(6):533-42.
PUBMED | CROSSREF

19. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. *Sci Transl Med* 2012;4(125):125ra31.
PUBMED | CROSSREF

20. Wang SV, Gagne JJ, Maro JC, Eworuke E, Kattinakere S, Kulldorff M, et al. Development and Evaluation of a Global Propensity Score for Data Mining with Tree-Based Scan Statistics. Sentinel; 2018.

21. Choi B, Kim SH, Lee H. Are registration of disease codes for adult anaphylaxis accurate in the emergency department? *Allergy Asthma Immunol Res* 2018;10(2):137-43.
PUBMED | **CROSSREF**

22. Helgeland J, Tomic O, Hansen TM, Kristoffersen DT, Hassani S, Lindahl AK. Postoperative wound dehiscence after laparotomy: a useful healthcare quality indicator? A cohort study based on Norwegian hospital administrative data. *BMJ Open* 2019;9(4):e026422.
PUBMED | **CROSSREF**

23. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, 1135-44.

24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *Journal of machine Learning research* 2011;12:2825-30.

25. Regateiro FS, Marques ML, Gomes ER. Drug-induced anaphylaxis: an update on epidemiology and risk factors. *Int Arch Allergy Immunol* 2020;181(7):481-7.
PUBMED | **CROSSREF**

26. Montañez MI, Mayorga C, Bogas G, Barrionuevo E, Fernandez-Santamaria R, Martin-Serrano A, et al. Epidemiology, mechanisms, and diagnosis of drug-induced anaphylaxis. *Front Immunol* 2017;8:614.
PUBMED | **CROSSREF**

27. Schweizer MT, Huang P, Kattan MW, Kibel AS, de Wit R, Sternberg CN, et al. Adjuvant leuprolide with or without docetaxel in patients with high-risk prostate cancer after radical prostatectomy (TAX-3501). *Cancer* 2013;119(20):3610-8.
PUBMED | **CROSSREF**

28. Tse SS, Kish T. Octreotide-associated neutropenia. *Pharmacotherapy* 2017;37(6):e32-7.
PUBMED | **CROSSREF**

29. Kim MJ, Shim DH, Cha HR, Kim CB, Kim SY, Park JH, et al. Delayed-onset anaphylaxis caused by IgE response to influenza vaccination. *Allergy Asthma Immunol Res* 2020;12(2):359-63.
PUBMED | **CROSSREF**

30. Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* 2018;19(1):270.
PUBMED | **CROSSREF**

31. FDA. Ketorolac tromethamine. https://www.accessdata.fda.gov/drugsatfda_docs/label/2014/074802s038lbl.pdf. Updated 2014. Accessed March 3, 2021.

32. Dhakal OP, Dhakal M, Bhandari D. Domperidone-induced dystonia: a rare and troublesome complication. *BMJ Case Rep* 2014;2014:bcr2013200282.
PUBMED | **CROSSREF**

33. Kaplan SA, Chughtai BI. Safety of tamsulosin: a systematic review of randomized trials with a focus on women and children. *Drug Saf* 2018;41(9):835-42.
PUBMED | **CROSSREF**