scientific reports

Check for updates

OPEN

Artificial intelligence models predicting abnormal uterine bleeding after COVID-19 vaccination

Yunjeong Choi[1,14], Jaeyu Park[2,14], Hyejun Kim[2,3,14], Young Joo Lee[4,14], Yongbin Lee[1,13], Yong Sung Choi[5], Seung Geun Yeo[6], Jiseung Kang[7,8], Masoud Rahmati[9,10,11], Hayeon Lee[1,2,12,15], Dong Keon Yon[1,2,5,15]✉ & Jinseok Lee[1,12,15]✉

The rapid deployment of COVID-19 vaccines has necessitated the ongoing surveillance of adverse events, with abnormal uterine bleeding (AUB) emerging as a reported concern in vaccinated females. We aimed to develop a machine learning (ML) model to predict post-vaccination AUB in women aged less than 50 years. A large-scale national cohort, the Korean Nationwide Cohort (K-COV-N cohort), was utilized, comprising over 7 million participants. The study employed advanced ML techniques, including ensemble models combining gradient boosting machine and logistic regression, and conducted feature importance analysis. The dataset was meticulously curated, focusing on relevant demographics and variables, and balanced using Synthetic Minority Over-sampling Technique. Using a national cohort of over 2 million COVID-19 vaccinated cases in South Korea, we developed a ML model for AUB prediction. Our study is the first to develop a predictive model for post-vaccination AUB, employing feature importance analysis to identify the key contributing factors. The analysis revealed three primary predictive features: COVID-19 vaccination frequency, NVX-CoV2373 (Novavax) COVID-19 vaccination count, and hemoglobin levels. These findings provide valuable insights into predicting the risk AUB following COVID-19 vaccination, potentially enhancing post-vaccination monitoring strategies.

Keywords COVID-19 vaccination, Abnormal uterine bleeding, Machine learning, Ensemble models

Abbreviations

AdaBoost	Adaptive boosting
AI	Artificial intelligence
AUB	Abnormal uterine bleeding

[1]Department of Biomedical Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do 17104, South Korea. [2]Center for Digital Health, Medical Science Research Institute Kyung Hee University Medical Center, Kyung Hee University College of Medicine, 23, Kyungheedae-ro, Dongdaemun-gu, Seoul 02447, Republic of Korea. [3]Department of Applied Information Engineering, Yonsei University, Seoul, South Korea. [4]Department of Obstetrics and Gynecology, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, South Korea. [5]Department of Pediatrics, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, South Korea. [6]Department of Otorhinolaryngology Head & Neck Surgery, Kyung Hee University School of Medicine, Kyung Hee University Medical Center, Seoul, South Korea. [7]School of Health and Environmental Science, College of Health Science, Korea University, Seoul, South Korea. [8]Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA. [9]Department of Physical Education and Sport Sciences, Faculty of Literature and Human Sciences, Lorestan University, Khorramabad, Iran. [10]Department of Physical Education and Sport Sciences, Faculty of Literature and Humanities, Vali-E-Asr University of Rafsanjan, Rafsanjan, Iran. [11]CEReSS-Health Service Research and Quality of Life Center, Assistance Publique-Hopitaux de Marseille, Aix-Marseille University, Marseille, France. [12]Department of Electronics and Information Convergence Engineering, Kyung Hee University, Yongin, South Korea. [13]Department of Biomedical Engineering, University of Connecticut, Storrs, CT, USA. [14]Yunjeong Choi, Jaeyu Park, Hyejun Kim and Young Joo Lee have contributed equally to this work. [15]Hayeon Lee, Dong Keon Yon and Jinseok Lee have contributed equally to this work. ✉email: yonkkang@gmail.com; gonasago@khu.ac.kr

AUROC	Area under receiver operating characteristic
GBM	Gradient boost model
KCDA	Korean Disease Control and Prevention Agency
LR	Logistic regression
ML	Machine learning
RF	Random forest
SD	Standard deviation
SMOTE	Synthetic minority oversampling technique
XGBoost	Extreme gradient boosting

The COVID-19 pandemic, which broke out in 2019, caused paralysis of medical profession and economic collapse worldwide[1]. In response, COVID-19 vaccines were rapidly developed through collaborative efforts between international organizations and companies[2]. The vaccines received emergency approval, bypassing some stages of clinical trials[3]. Since the vaccines were released before completing all trials, there were uncertainties regarding their safety, and diverse side effects were reported after their release[4]. The most common side effects included fever, fatigue, and headache[5,6], numerous reports of abnormal uterine bleeding in women[7]. At this time, the ability to predict such adverse effects prior to vaccination could greatly enhance recipient safety and bolster vaccine confidence. This study developed a machine learning (ML) model to predict post-vaccination side effects, particularly focusing on abnormal uterine bleeding (AUB). This model has the potential to monitor adverse events and enhance vaccine safety. However, its feasibility and effectiveness require further validation. Thus, this study did not only on developing this ML model, but also rigorously evaluated its practicality for predicting post-vaccination side effects, especially AUB.

Previous studies have reported varying rates of menstrual disorders following COVID-19 vaccination, with approximately 0.5–2% of females experiencing such effects[8]. However, no studies to date have attempted to predict AUB following vaccination. Several studies have developed prediction models for common COVID-19 vaccine adverse effects. One study using vaccine adverse event reporting system data (40,000 participants) achieved 75% accuracy[9], another from Israel (20,000 participants) achieved 69–74% accuracy for effects like weakness and fever[10]. The other study using data from Iran and Switzerland (19,000 participants) achieved 62–69% accuracy[17]. Thus, utilizing feature importance analysis, our study aimed to develop an ML model specifically for predicting AUB following vaccination.

Methods

Data source

We used national, grand-scale, and general population cohort studies, including: a nationwide Korean cohort (K-COV-N cohort; total $n = 7,281,607$). This study was approved by the Institutional Review Boards of the Korean Disease Control and Prevention Agency (KDCA) and Kyung Hee University (KHUH 2022-06-042). Because this study used deidentified national data, informed consent was waived by KDCA and Kyung Hee University. All processes were conducted in accordance with the relevant regulations.

Data selection

The K-COV-N cohort was built and provided by the National Health Insurance Service of South Korea and KDCA based on participants aged ≥ 20 years between January 1, 2018 and December 31, 2022. In the K-COV-N cohort, a comprehensive analysis involving multi-year examinations of each subject was conducted, resulting in a dataset of 20,343,949 entries. We applied the exclusion criteria based on the following: (1) participants who are male; (2) aged under 50 years; or (3) those with insufficient information[11]. The exclusion criteria were determined considering that AUB primarily affects individuals of reproductive age[11,12]. The selected data consisted of 5,304,203 cases, from which 2,296,675 cases of those who received the COVID-19 vaccine were further refined (Fig. 1). AUB was defined as a side effect occurring between the time of vaccination and the next menstrual cycle[12], based on research findings showing that such symptoms predominantly appear within this period[12].

Statistical analysis

We investigated the risk of AUB after the COVID-19 vaccination using data from 5,304,203 cases, including both vaccinated and unvaccinated individuals. Primary exposure was defined as COVID-19 vaccination status. A multivariable logistic regression model was used to derive propensity scores by adjusting for related covariates[13,14]. Using the calculated overlap weights, a propensity score-based overlap-weighted cohort was constructed[13]. For unvaccinated individuals, the individual index date was randomly assigned to ensure a uniform distribution of vaccination dates among vaccinated participants[13–16]. The outcome of interest, AUB, was defined based on ICD-10 codes as a side effect incidence between the time of vaccination and the next menstrual cycle[12]. We derived the adjusted hazard ratios (aHRs) and the corresponding 95% confidence intervals (CIs) through the Cox proportional hazards regression model.

ML model development and validation

To train and validate our ML model for the prediction of AUB occurrence after COVID-19 vaccination, we split the K-COV-N dataset ($n = 2,296,394$) into training ($n = 1,837,115$) and test ($n = 459,279$) datasets in a ratio of 8:2 in a stratified manner. Using the training set ($n = 1,837,115$), we conducted 5-fold cross-validation to verify the generalizability of the generated ML model[17]. Thereafter, we tested the model with separate test data ($n = 459,279$). This test set was exclusively used for the independent evaluation of the model, not for its training or internal validation[18].

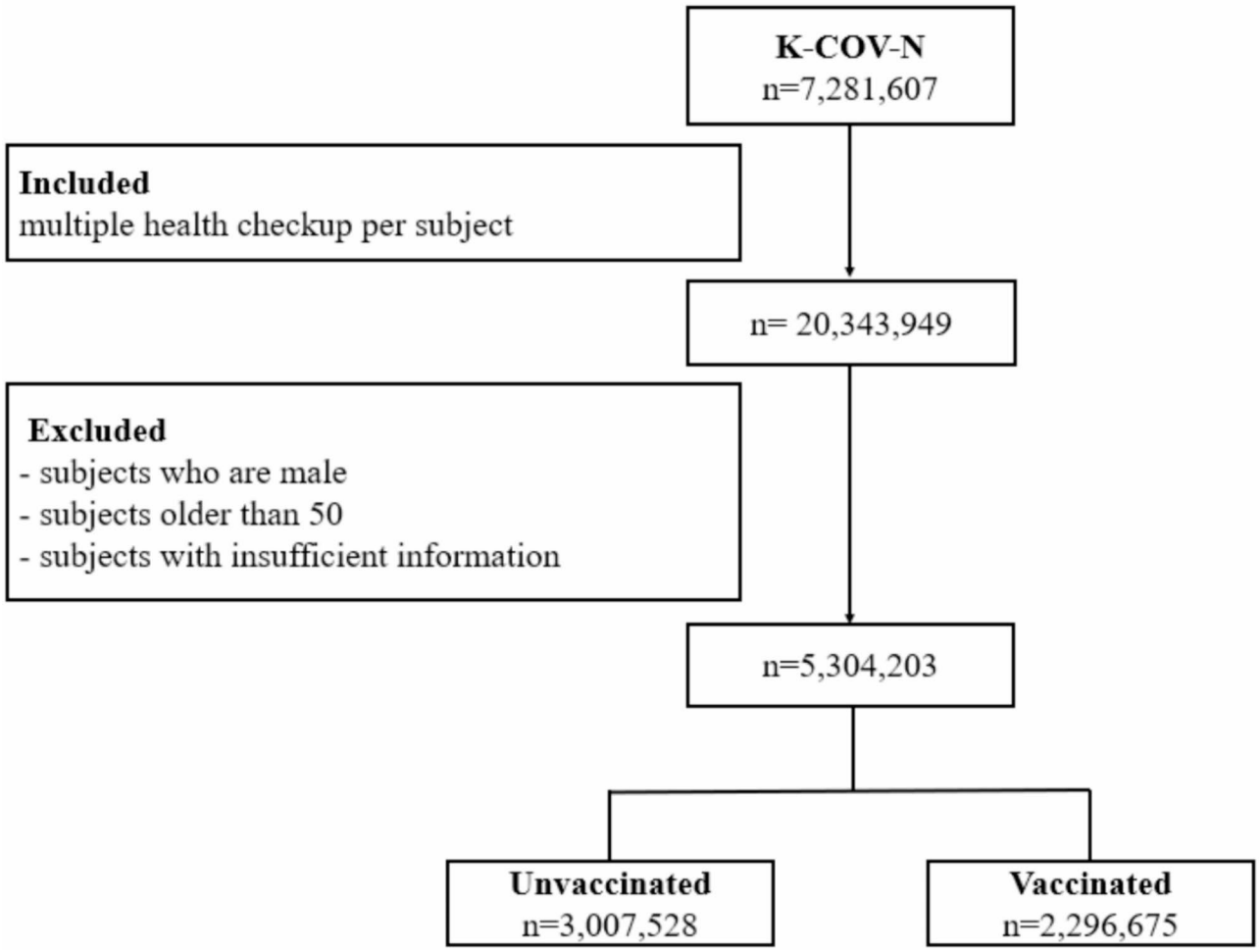

Fig. 1. Study population and data selection process from K-COV-N.

For the ML model input, we used the following 46 variables provided by K-COV-N: age, household income, region of residence, body mass index (BMI), systolic blood pressure, diastolic blood pressure, fasting blood glucose, serum total cholesterol, hemoglobin, serum glutamic oxaloacetic transaminase, serum glutamic pyruvate transaminase, epidermal growth factor receptor, smoking status, alcoholic consumption, physical activity sessions, history of chronic kidney diseases, chronic infectious diseases, chronic liver diseases, blood diseases, musculoskeletal diseases, rhinitis, asthma, food allergy, atopy, chronic obstructive pulmonary disease, hypertension, type 2 diabetes, cancer, rheumatisms, dyslipidemia, menstrual irregularity, amenorrhea, cramps, previous use of inhaled steroids, intranasal steroids, topical steroids, immunosuppressants, oral steroids, type of vaccine, the number of total vaccinations, ChAdOx1-S (AstraZeneca) vaccinations, BNT162b2 (Pfizer/BioNTech) vaccinations, mRNA-1273 (Moderna) vaccinations, Ad26.COV2.S (Johnson & Johnson–Janssen) vaccinations, and NVX-CoV2373 (Novavax) vaccinations.

Specifically, household income included four percentile ranges: 0–25, 26–50, 51–75, and 76–100. The residential regions were divided into two categories: rural and urban. Smoking status was classified as yes or no. Alcoholic consumption was classified into four categories: less than once, 1–2, 3–4, and 5 or more days/week. Physical activity included eight categories: exercise: 0, 1, 2, 3, 4, 5, 6, and 7 days/week. History of diseases and previous use of medications were categorized as either yes or no. The vaccine types included five categories: ChAdOx1-S, BNT162b2, mRNA-1273, Ad26.COV2.S, and NVX-CoV2373. The total number of vaccinations was categorized into four groups: 1, 2, 3, and 4 doses, and the number of vaccinations for each vaccine type was also categorized into four groups: 0, 1, 2, and 3 doses.

Data imbalance

There was a significant imbalance in the data, with only 1719 (0.075%) AUB cases compared to 2,294,675 (99.925%) non-AUB cases. To minimize the model bias towards a majority, the non-AUB group, we balanced the two groups using the synthetic minority oversampling technique (SMOTE)[19,20] by up-sampling the AUB data. Subsequently, we identified demographic parity-based coefficients for the weighted value search[21,22] in an iterative manner, yielding a closed-form expression for the data weights. These methods effectively mitigated the bias towards the non-AUB data.

Model selection and performance assessment

To predict AUB occurrence after COVID-19 vaccination, we explored five ML models: adaptive boosting (AdaBoost)[23], extreme gradient boosting (XGBoost)[24], gradient boosting machine (GBM)[25], random forest (RF)[26], and logistic regression (LR)[27]. Five predictive models were initially developed and evaluated using cross-validation to assess their individual performance. The top three models, as determined by their cross-validation metrics, were further analyzed for ensemble configurations. All possible combinations of the top three models were tested, and their performances were compared[28]. Ultimately, we identified the most effective ML model for predicting post-vaccination AUB.

Performance evaluations utilized five-fold cross-validation across five metrics: sensitivity, specificity, accuracy, balanced accuracy, and area under the receiver operating characteristic (AUROC)[29]. For each model type, five models were created through 5-fold cross-validation. The average and standard deviation (SD) of the performance indicators of these five models were used as the standard for evaluating them. Owing to significant data imbalance, balanced accuracy and AUROC were prioritized as the primary metrics for evaluating the model performance[30].

We present a feature importance analysis from our model, listing features in the order they contributed to the AUB prediction. In decision tree approaches such as AdaBoost, XGBoost, GBM and RF, the calculation of feature importance values relies on evaluating the reduction in node impurity, considering the likelihood of reaching each node. For the node impurity, we used the Gini index, a measure that quantifies the level of impurity at a particular node by assessing the extent to which a specific variable would be inaccurately classified if chosen randomly. Impurity is assessed by factoring in the weighted sum of the squared probabilities of each class within the node. The comprehensive approach integrates node probabilities, impurity metrics, and feature importance computations, providing a nuanced understanding of the diverse influences of various variables on prediction outcomes ("Supplementary Methods"). To address the risk of Type I error due to multiple comparisons, we applied a Bonferroni correction to the feature importance analysis and the statistical comparisons in Table S2. Given the 46 features included in the model, the α-threshold was adjusted to 0.0011 (0.05/46)[31]. Features and variables that did not meet this corrected threshold were not considered statistically significant. This correction ensures a robust and conservative interpretation of the results.

Software and libraries

The models were developed in Python (version: 3.5.2) using NumPy (version: 1.16.2), Pandas (version: 0.24.2), Matplotlib (version: 3.0.3), and Scikit-learn (version: 0.20.3) software. All statistical analyses were performed using the SAS (version 9.4; SAS Institute Inc., Cary, NC, USA) and R software (version 3.1.1; R Foundation, Vienna, Austria)[22].

Results

Statistical analysis

Table S1 presents the baseline characteristics of the overlap-weighted cohort in South Korea. Compared to the non-vaccinated control group, individuals with vaccination had a significantly higher risk of AUB (aHR, 1.21 [95% CI, 1.13–1.29]) (Table 1). This association was more pronounced among females aged 35–50 years (aHR, 1.28 [95% CI, 1.18–1.39]). Stratified analyses by region of residence, household income, and BMI revealed consistent findings, except among individuals with high household income, where the association was less pronounced. Detailed results for stratification are presented in Table 1.

Baseline characteristics

Table S2 shows a comparison of the variables included in the K-COV-N to predict occurrence of AUB after COVID-19 vaccination based on regular health checkups. The patients with AUB showed a mean age of 44.7 (SD: 2.8), whereas individuals without AUB showed a mean age of 44.6 (SD: 2.9).

Five-fold cross validation results

The following hyperparameters were utilized for the five ML model. XGBoost was implemented with a column subsample by tree 1, learning rate of 0.3, maximum depth of 6, and number of estimators 100. The RF parameters consisted of no maximum depth, auto maximum features, minimum samples per leaf of 1, minimum samples per split of 2, number of estimators 100, and balanced class weight. The GBM utilized a learning rate of 0.1, maximum depth of 3, minimum samples per leaf of 1, minimum samples per split of 2, and number of estimators 100. AdaBoost was configured with a stagewise additive modeling algorithm using a multiclass exponential loss function, real variant; 50 tree estimators and a learning rate of 1. The LR parameters included the solver of the library for large linear classification; the penalty norm was L2, the inverse of the regularization strength was 1, and the maximum number of iterations was 100. To enhance model performance, the optimal thresholds were determined for individual models. These threshold values were established as follows: 0.492 for AdaBoost, 0.35 for XGBoost, 0.16 for GBM, 0.1 for RF, and 0.492 for LR. For the ensemble models, the hyperparameters remained consistent with those used in the single ML models. The optimization process revealed the following weight combinations: AdaBoost and GBM models with weights of 1:1, AdaBoost and LR models with weights of 2:3, GBM and LR models with weights of 2:3, and AdaBoost, GBM, and LR models with weights of 2:2:3.

The 5-fold cross-validation results are presented in Table 2. All 5-fold cross-validation results were expressed as the average and SD of the five models for each model type. Among the single models, LR, AdaBoost and GBM provided the best performance based on balanced accuracy and AUROC. The LR model achieved a mean balanced accuracy of 60.9% (SD: 0.006) and a mean AUROC of 0.657 (SD: 0.005). The AdaBoost model showed a mean balanced accuracy of 57.7% (SD: 0.014) and a mean AUROC of 0.626 (SD: 0.014). Finally, the GBM model yielded a mean balanced accuracy of 57% (SD: 0.009) and a mean AUROC of 0.630 (SD: 0.005). To

Outcome	Event, *n* (%)	aHR* (95% CI)
Abnormal uterine bleeding		
Non-vaccination control	1519 (1.19)	1.00 (reference)
Individuals with vaccination	1828 (1.43)	**1.21 (1.13 to 1.29)**
Age		
20–34 y		
Non-vaccination control	1033 (1.71)	1.00 (reference)
Individuals with vaccination	1292 (2.14)	1.06 (0.93 to 1.20)
35–50 y		
Non-vaccination control	486 (0.72)	1.00 (reference)
Individuals with vaccination	537 (0.79)	**1.28 (1.18 to 1.39)**
Region of residence		
Rural		
Non-vaccination control	735 (1.23)	1.00 (reference)
Individuals with vaccination	874 (1.46)	**1.19 (1.08 to 1.31)**
Urban		
Non-vaccination control	784 (1.15)	1.00 (reference)
Individuals with vaccination	954 (1.40)	**1.23 (1.12 to 1.35)**
Household income		
Low (0–39 percentiles)		
Non-vaccination control	571 (1.21)	1.00 (reference)
Individuals with vaccination	681 (1.44)	**1.19 (1.06 to 1.33)**
Middle (40–79 percentiles)		
Non-vaccination control	542 (1.08)	1.00 (reference)
Individuals with vaccination	698 (1.40)	**1.30 (1.16 to 1.45)**
High (80–100 percentiles)		
Non-vaccination control	406 (1.32)	1.00 (reference)
Individuals with vaccination	450 (1.46)	1.12 (0.98 to 1.29)
BMI		
< 23.0 kg/m^2		
Non-vaccination control	803 (1.31)	1.00 (reference)
Individuals with vaccination	991 (1.61)	**1.24 (1.13 to 1.36)**
23.0–24.9 kg/m^2		
Non-vaccination control	319 (1.15)	1.00 (reference)
Individuals with vaccination	370 (1.34)	**1.17 (1.01 to 1.36)**
≥ 25.0 kg/m^2		
Non-vaccination control	397 (1.02)	1.00 (reference)
Individuals with vaccination	468 (1.20)	**1.18 (1.03 to 1.35)**
Smoking status		
Non-smoker		
Non-vaccination control	1500 (1.19)	1.00 (reference)
Individuals with vaccination	1801 (1.43)	**1.21 (1.13 to 1.29)**
Smoker		
Non-vaccination control	19 (1.01)	1.00 (reference)
Individuals with vaccination	28 (1.47)	1.44 (0.80 to 2.57)
Alcohol consumption		
Non-drinker		
Non-vaccination control	1330 (1.16)	1.00 (reference)
Individuals with vaccination	1590 (1.39)	**1.20 (1.11 to 1.29)**
Drinker		
Non-vaccination control	190 (1.43)	1.00 (reference)
Individuals with vaccination	239 (1.80)	**1.30 (1.07 to 1.57)**
Aerobic physical activity		
Insufficient physical activity		
Non-vaccination control	913 (1.21)	1.00 (reference)
Individuals with vaccination	1072 (1.43)	**1.18 (1.08 to 1.29)**
Sufficient physical activity		
Continued		

Outcome	Event, *n* (%)	aHR* (95% CI)
Non-vaccination control	607 (1.15)	1.00 (reference)
Individuals with vaccination	757 (1.43)	**1.25 (1.12 to 1.39)**
Charlson comorbidity index		
0 score		
Non-vaccination control	1268 (1.15)	1.00 (reference)
Individuals with vaccination	1572 (1.43)	**1.27 (1.18 to 1.36)**
≥1 score		
Non-vaccination control	251 (1.39)	1.00 (reference)
Individuals with vaccination	257 (1.42)	0.94 (0.79 to 1.12)

Table 1. Overlap-weighted HR (95% CI) for the risk of AUB event following COVID-19 vaccination. aHR, adjusted hazard ratio; AUB, abnormal uterine bleeding; BMI, body mass index; CI, confidence interval. The data in bold indicate significant differences (p-value < 0.05). ***aHR**: adjusted for age (20–39, 40–59, and ≥ 60 years); sex (male and female); region of residence (urban and rural); household income (low [0–39 quartiles], middle [40–79 quartiles], and high [80–100 quartiles]); Charlson comorbidity index score (0, 1, and ≥ 2); body mass index (underweight [< 18.5 kg/m^2], normal weight [18.5–22.9 kg/m^2], overweight [23.0–24.9 kg/m^2], and obesity [≥ 25.0 kg/m^2]); blood pressure ([systolic blood pressure < 140 mmHg and diastolic blood pressure < 90 mmHg] and [systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg]); fasting blood glucose (< 100 and ≥ 100 mg/dL); glomerular filtration rate (< 60, 60–89, and ≥ 90 mL/min/1.73 m^2); smoking status (never and [ex- or current smoker]); alcohol consumption (drinks; <1, 1–2, 3–4, and ≥ 5 days per week); aerobic physical activity (sufficient and insufficient); and history of medication use for diabetes mellitus, hyperlipidemia, and hypertension.

Model	Sensitivity, mean (SD)	Specificity, mean (SD)	Accuracy, mean (SD)	Balanced accuracy, mean (SD)	AUROC, mean (SD)
AdaBoost	0.542	0.612	0.612	0.577	0.626
	(0.047)	(0.027)	(0.027)	(0.014)	(0.014)
XGBoost	0.065	0.961	0.960	0.513	0.635
	(0.017)	(0.007)	(0.007)	(0.005)	(0.007)
GBM	0.508	0.632	0.632	0.570	0.630
	(0.036)	(0.025)	(0.025)	(0.009)	(0.005)
RF	0.060	0.968	0.967	0.514	0.514
	(0.016)	(0.001)	(0.001)	(0.008)	(0.008)
LR	0.622	0.596	0.596	0.609	0.657
	(0.011)	(0.002)	(0.002)	(0.006)	(0.005)
AdaBoost + GBM	0.630	0.535	0.535	0.582	0.630
	(0.037)	(0.029)	(0.029)	(0.009)	(0.006)
AdaBoost + LR	0.622	0.596	0.596	0.609	0.658
	(0.011)	(0.002)	(0.002)	(0.006)	(0.005)
GBM + LR	**0.622**	**0.596**	**0.596**	**0.609**	**0.659**
	(0.011)	**(0.002)**	**(0.002)**	**(0.006)**	**(0.006)**
AdaBoost + GBM + LR	0.549	0.617	0.616	0.583	0.659
	(0.040)	(0.024)	(0.024)	(0.011)	(0.006)

Table 2. Comparison results of the 5-fold cross-validation. AdaBoost, adaptive boosting; AUROC, area under receiver operating characteristic; GBM, gradient boost model; LR, logistic regression; RF, random forest; SD, standard deviation; XGBoost, extreme gradient boosting. Bold indicates machine learning model with best performance of prediction.

enhance the prediction accuracy, we examined an ensemble approach by combining the top-3 single models: LR, AdaBoost, and GBM. This approach generally yields higher accuracy than individual models. Notably, the combination of the GBM and LR outperformed all other single or combined models by achieving a balanced accuracy of 60.9% (SD: 0.006) and an AUROC of 0.659 (SD: 0.006).

Feature importance analysis

Feature importance analysis was performed to analyze the effect of each feature on predicting the occurrence of AUB. Figure 2 shows the ranked normalized feature importance from the ensemble model combining the GBM and LR. Feature importance was first calculated separately: GBM importance was based on reduction in loss

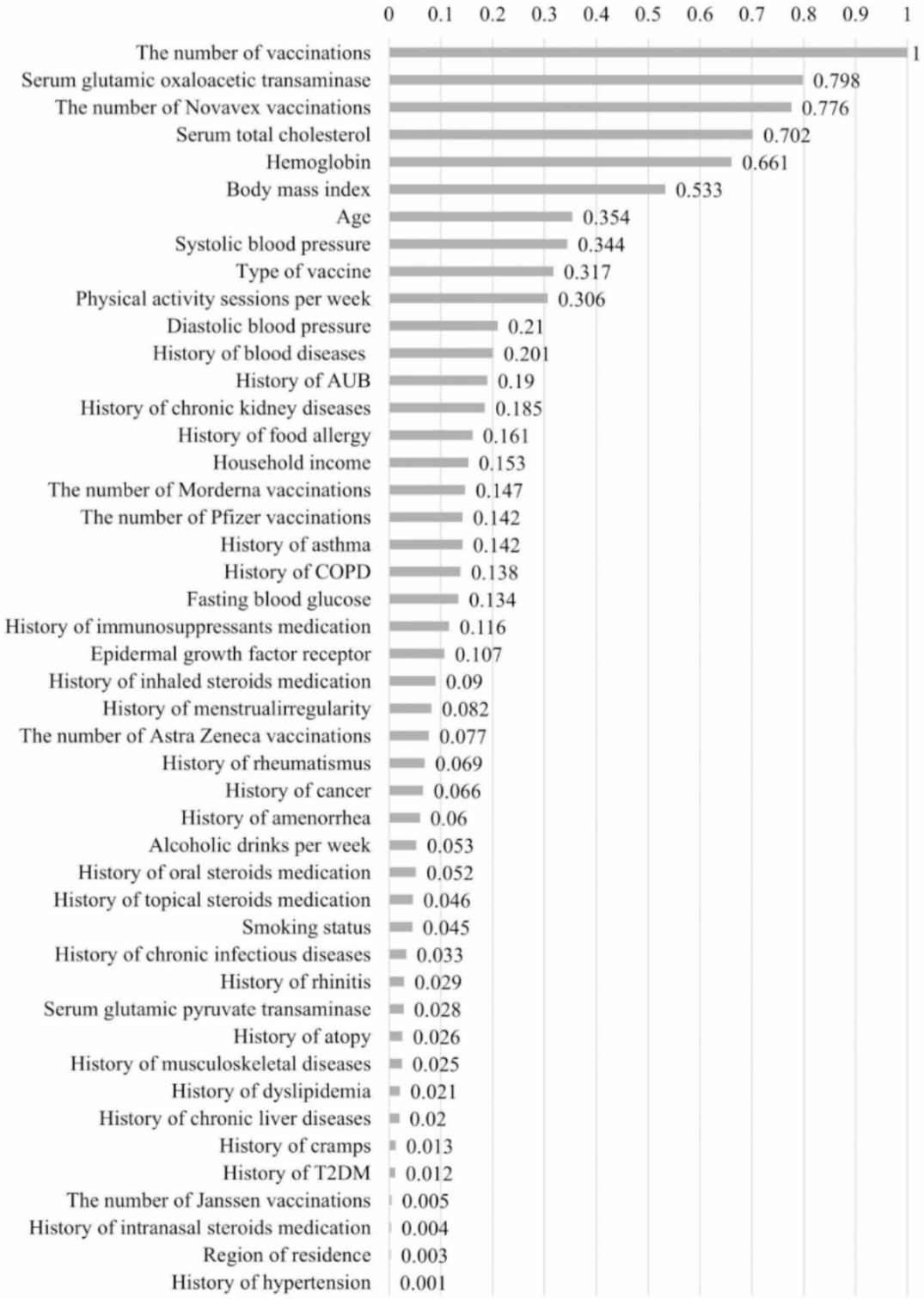

Fig. 2. Results of the ranked normalized feature importance from the ensemble model combining GBM and LR. *Abbreviations*: AUB, Abnormal uterine bleeding; COPD, chronic obstructive pulmonary disease; GBM, gradient boosting machine; LR, logistic regression T2DM, type 2 diabetes mellitus.

during tree splits, and LR importance was derived from the magnitude of standardized regression coefficients. These values were normalized by setting the most important feature in each model to 1 and scaling others relative to it. The normalized importance values were then averaged across the two models to produce the final rankings.

Model	Sensitivity	Specificity	Accuracy	Balanced accuracy	AUROC
AdaBoost	0.549	0.591	0.591	0.570	0.617
XGBoost	0.544	0.591	0.591	0.567	0.594
GBM	0.552	0.614	0.614	0.583	0.629
RF	0.047	0.970	0.969	0.508	0.508
LR	0.640	0.594	0.594	0.617	0.661
AdaBoost + GBM	0.672	0.514	0.514	0.593	0.630
AdaBoost + LR	0.640	0.594	0.594	0.617	0.662
GBM + LR	**0.640**	**0.594**	**0.594**	**0.617**	**0.661**
AdaBoost + GBM + LR	0.602	0.594	0.594	0.598	0.661

Table 3. Comparison of the AUB prediction performances of the machine learning models on the test data set. AdaBoost, adaptive boosting; AUB, Abnormal uterine bleeding; AUROC, area under receiver operating characteristic; GBM, gradient boost model; LR, logistic regression; RF, random forest; XGBoost, extreme gradient boosting. Bold indicates machine learning model with best performance of prediction.

According to the results, the number of vaccinations had the highest importance value among the features, followed by serum glutamic oxaloacetic transaminase, the number of Novavax vaccinations, serum total cholesterol and hemoglobin. Feature importances are as follows: the number of vaccinations, 1.00; serum glutamic oxaloacetic transaminase, 0.798; the number of Novavax vaccinations, 0.776; serum total cholesterol, 0.702; hemoglobin, 0.661; BMI, 0.533; age, 0.354; systolic blood pressure, 0.344; type of vaccine, 0.317; physical activity sessions per week, 0.306; diastolic blood pressure, 0.21; history of blood diseases, 0.201; history of AUB, 0.19; history of chronic kidney diseases, 0.185; history of food allergy, 0.161; household income, 0.153; the number of Moderna vaccinations, 0.147; the number of Pfizer vaccinations, 0.142; history of asthma, 0.142; history of COPD, 0.138; fasting blood glucose, 0.134; history of immunosuppressants medication, 0.116; epidermal growth factor receptor, 0.107; history of inhaled steroids medication, 0.09; history of menstrual irregularity, 0.082; the number of Astra Zeneca vaccinations, 0.077; history of rheumatismus, 0.069; history of cancer, 0.066; history of amenorrhea, 0.06; alcoholic drinks per week, 0.053; history of oral steroids medication, 0.052; history of topical steroids medication, 0.046; smoking status, 0.045; history of chronic infectious diseases, 0.033; history of rhinitis, 0.029; serum glutamic pyruvate transaminase, 0.028; history of atopy, 0.026; history of chronic liver diseases, 0.02; history of cramps, 0.013; history of T2DM, 0.012; the number of Janssen vaccinations, 0.005; history of intranasal steroids medication, 0.004; region of residence, 0.003; and history of hypertension, 0.001.

Test data results

Table 3 summarizes the test data results. The test data was isolated dataset ($n = 459{,}279$) from the K-COV-N. Similar to 5-fold cross-validation, the test results showed that the ensemble model combining the GBM and LR showed the highest performance. The balance accuracy of the model was 61.7% and the AUROC was 0.661. The similarity between the results of the cross-validation and test data indicated that overfitting or underfitting was minimal in the model.

Discussion

Utilizing a comprehensive, large-scale national cohort of more than 2 million participants who received COVID-19 vaccination in South Korea, this study developed an ML model to predict AUB. This research has several notable strengths. First, to our knowledge, this is the first study to develop a ML model for AUB prediction. While several cohort studies conducted during the COVID-19 pandemic have revealed associations between AUB and COVID-19 vaccination, none have previously developed a predictive ML model. Second, through feature importance analysis, we identified the top three features contributing to AUB: the number of vaccinations, the number of Novavax vaccinations, and hemoglobin concentration. Finally, our additional analysis of the association between COVID-19 vaccination and AUB supports our hypothesis that vaccinated individuals are more likely to experience AUB, emphasizing the importance of rigorous statistical methods in vaccine safety monitoring. Future studies should consider advanced methods, such as regularization or dimensionality reduction, to address collinearity and improve the precision of effect estimates. By integrating predictive modeling with robust statistical analyses, we aim to enhance the generalizability of our findings and contribute to the ongoing evaluation of vaccine safety.

The association between vaccination and AUB has been previously reported for other vaccines, including human papillomavirus and hepatitis B vaccines, although the precise mechanisms remain incompletely understood[32]. Several hypotheses have also been proposed to explain vaccination-related AUB. One key mechanism involves the disruption of the hypothalamic-pituitary-ovarian axis, which regulates the menstrual cycle[33]. This axis is susceptible to external factors such as stress, environmental changes, and hormone therapy, suggesting that vaccination could serve as a triggering factor. The systemic inflammatory response following vaccination may disrupt this axis, potentially leading to AUB[34]. Additionally, vaccination-associated immune-mediated thrombocytopenia could affect normal coagulation pathways in the endometrium, potentially compromising repair mechanisms and resulting in heavy bleeding[35,36]. Our findings highlight several important factors contributing to post-vaccination AUB. The immune response triggered by vaccines, particularly elevated cytokine levels, can disrupt hormonal balance and affect menstrual cycles[37,38]. Notably, our study aligns with

previous research showing that individuals who received multiple vaccinations had a higher likelihood of experiencing menstrual irregularities[33], suggesting a potential dose-dependent effect. This implies that frequent vaccinations may amplify these association. These interconnected factors provide a comprehensive framework for understanding the development of post-vaccination AUB.

In this study, several limitations were identified. First, given the model's performance and the exclusive use of the Korean cohort (K-COV-N) without external validation, there may be limitations in the generalizability of our findings globally. To address this, the study employed rigorous data curation and analysis methods to ensure that the results were as representative and comprehensive as possible within the given cohort[39]. Second, the disproportionate representation of AUB cases in the dataset, particularly the low incidence of AUB cases compared to non-AUB cases, posed a significant challenge to the ML model's accuracy and generalizability. We acknowledge that the balanced accuracy of 60.9% (SD: 0.006) and AUROC of 0.659 (SD: 0.006) for cross-validation results, along with the balanced accuracy of 61.7% and AUROC of 0.661 on the test dataset, indicate moderate predictive performance. These results are primarily influenced by the challenges of predicting rare outcomes, such AUB, which constitutes only 0.075% of the dataset. Although we applied SMOTE to address the data imbalance[40,41], the rarity of the event posed challenges to achieving higher predictive accuracy. This limitation underscores the inherent difficulty in developing robust predictive models for rare outcomes. Future research should focus on incorporating additional potential risk factors, such as hormonal levels, genetic predispositions, and other clinical variables, to enhance model performance. Furthermore, external validation using independent cohorts from diverse populations is essential to evaluate the generalizability and robustness of the model. Advanced methods, such as feature selection optimization and regularization techniques, could also be explored to improve predictive accuracy and establish clinically reliable models. Third, statistical sensitivity analyses were conducted to evaluate the robustness of our findings. To mitigate the risk of Type I error associated with multiple comparisons, we applied a Bonferroni correction to the feature importance analysis and statistical comparisons, including those in Table S2. This adjustment, with a corrected α-threshold of 0.0011, minimizes the potential for false positives[31]. While this approach reduces Type I error, we acknowledge its conservative nature and the possibility of overlooking moderate associations. Additionally, potential collinearity and confounding factors among the variables were considered. Future research could explore advanced statistical methods, such as regularization techniques or dimensionality reduction, to address collinearity and identify meaningful associations that might be missed due to conservative corrections. These approaches could further validate and refine the predictive capabilities of our model, enhancing its generalizability and utility in clinical settings. Finally, there is a potential surveillance bias in AUB diagnosis. While our study validated the association between COVID-19 vaccination and AUB, diagnostic rates may be influenced by the increased medical surveillance of vaccine recipients. Previous studies reporting elevated AUB risk post-vaccination have suggested that the observed rates could be partially attributed to heightened monitoring, particularly given the concurrent observation of bleeding events and thrombocytopenia[42]. This surveillance bias may affect both the detection and reporting of AUB cases, potentially affecting the accuracy of predictive capabilities of our ML model. The challenge of distinguishing between true incidence and surveillance-induced detection rates represents an important consideration in interpreting our findings and may have influenced our model's predictive performance.

Despite these limitations, this study has several strengths. As a pioneering large-scale national cohort investigation, this study sheds light on the relationship between COVID-19 vaccine side effects and AUB, thereby contributing significantly to vaccine safety and women's health knowledge. The meticulous use of a comprehensive dataset, particularly targeting women aged less than 50 years, enhances the integrity and relevance of the study's data. The implementation of an advanced ensemble model combining the GBM and LR presents exceptional predictive accuracy. Additionally, the study's thorough feature importance analysis and consistent results across cross-validation and test data affirm the model's reliability, effectively addressing concerns about overfitting or underfitting.

While COVID-19 vaccination may be associated with certain adverse effects, including AUB, its public health benefits of vaccination during the pandemic significantly outweigh these risks. Studies have indicated that vaccination-related AUB and other menstrual disorders are typically self-limiting and resolve spontaneously[43]. Nevertheless, given public concerns about post-vaccination AUB, despite its low absolute risk, vigilant monitoring of high-risk individuals remains crucial. These findings have important public health implications. The identification of key risk factors for post-vaccination AUB can inform targeted surveillance strategies, particularly for women under 50 years of age[44]. These findings suggest the need for enhanced monitoring protocols among healthcare providers, who should be adequately informed about these risk factors to ensure comprehensive documentation and reporting of AUB cases[45]. Implementation of these targeted monitoring and communication approaches could enhance the management of vaccine-related effects and, consequently, strengthen public trust in vaccination programs.

In conclusion, using a comprehensive national cohort of over 2 million COVID-19 vaccinated participants in South Korea, we developed and validated an ML model for predicting post-vaccination AUB. Our study represents the first attempt to create a predictive model specifically focused on post-vaccination AUB, employing robust feature importance analysis to identify key contributing factors. Through this analysis, we identified three primary predictive features that significantly influenced the occurrence of AUB: frequency of COVID-19 vaccinations, the number of Novavax vaccinations received, and hemoglobin concentration. These findings not only advance our understanding of post-vaccination AUB risk factors but also provide a foundation for developing more targeted monitoring strategies in clinical settings.

Data availability

Data are available on reasonable request. Study protocol, statistical code: available from DKY (email: yonkkang@gmail.com). Data set: available from the National Health Insurance Service (NHIS) in South Korea through a data use agreement.

Received: 7 February 2024; Accepted: 24 February 2025
Published online: 27 February 2025

References

1. Cha, Y., Jung, W., Seo, M. & Rahmati, M. The emerging pandemic recent: SARS-CoV-2. *Life Cycle* **3**, e2. https://doi.org/10.54724/lc.2023.e2 (2023).
2. Le, T. T. et al. The COVID-19 vaccine development landscape. *Nat. Rev. Drug Discov.* **19**, 305–306 (2020).
3. Vasireddy, D., Atluri, P., Malayala, S. V., Vanaparthy, R. & Mohan, G. Review of COVID-19 vaccines approved in the United States of America for emergency use. *J. Clin. Med. Res.* **13**, 204 (2021).
4. Gee, J. et al. First month of COVID-19 vaccine safety monitoring—United States, December 14, 2020–January 13, 2021. *Morb. Mortal. Wkly. Rep.* **70**, 283 (2021).
5. Kim, Y. et al. Shifts in COVID-19 vaccine acceptance rates among African countries: A systematic review and meta-analysis. *Life Cycle* **3**, e7. https://doi.org/10.54724/lc.2023.e7 (2023).
6. Lee, S. et al. Risks of chronic urticaria after SARS-CoV-2 infection: Binational population-based cohort studies from South Korea and Japan. *J. Allergy Clin. Immunol. Pract.* https://doi.org/10.1016/j.jaip.2024.05.044 (2024).
7. Nazir, M. et al. Menstrual abnormalities after COVID-19 vaccines: A systematic review. *Vacunas* **23**, S77–s87 (2022).
8. Paik, H. & Kim, S. K. Female reproduction and abnormal uterine bleeding after COVID-19 vaccination. *Clin. Exp. Reprod. Med.* **50**, 69–77. https://doi.org/10.5653/cerm.2023.05925 (2023).
9. Monadhel, H., Abbas, A. R. & Mohammed, A. J. COVID-19 vaccine: Predicting vaccine types and assessing mortality risk through ensemble learning algorithms. *F1000Research* **12**, 1200. https://doi.org/10.12688/f1000research.140395.1 (2023).
10. Kara, A. et al. Analysis of participant-reported adverse events following the first dose of inactivated SARS-Cov-2 vaccine (TURKOVAC™) through telephone survey in Türkiye. *Ann. Med.* **55**, 1070–1079. https://doi.org/10.1080/07853890.2023.2183985 (2023).
11. Albers, J. R., Hull, S. K. & Wesley, R. M. Abnormal uterine bleeding. *Am. Fam. Phys.* **69**, 1915–1926 (2004).
12. Issakov, G., Tzur, Y., Friedman, T. & Tzur, T. Abnormal uterine bleeding among COVID-19 vaccinated and recovered women: A national survey. *Reprod. Sci.* **30**, 713–721 (2023).
13. Kim, H. et al. Risks of cutaneous immune-related adverse events in long COVID: Multinational cohort studies in South Korea, Japan, and the UK. *J. Med. Virol.* **96**, e29740. https://doi.org/10.1002/jmv.29740 (2024).
14. Xie, Y., Choi, T. & Al-Aly, Z. Postacute sequelae of SARS-CoV-2 infection in the pre-delta, delta, and omicron eras. *N. Engl. J. Med.* **391**, 515–525. https://doi.org/10.1056/NEJMoa2403211 (2024).
15. Cai, M., Xie, Y., Topol, E. J. & Al-Aly, Z. Three-year outcomes of post-acute sequelae of COVID-19. *Nat. Med.* **30**, 1564–1573. https://doi.org/10.1038/s41591-024-02987-8 (2024).
16. Lee, K. et al. Long-term Gastrointestinal and hepatobiliary outcomes of COVID-19: A multinational population-based cohort study from South Korea, Japan, and the UK. *Clin. Mol. Hepatol.* **30**, 943–958. https://doi.org/10.3350/cmh.2024.0203 (2024).
17. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* **21**, 137–146. https://doi.org/10.1007/s11222-009-9153-8 (2011).
18. Ni, Z. Reframe the field of aerospace engineering via machine learning: Application and comparison. *J. Phys. Conf. Ser.* **2386**, 012031. https://doi.org/10.1088/1742-6596/2386/1/012031 (2022).
19. Batista, G. E., Prati, R. C. & Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **6**, 20–29 (2004).
20. Kim, J. et al. Quantification of identifying cognitive impairment using olfactory-stimulated functional near-infrared spectroscopy with machine learning: A post hoc analysis of a diagnostic trial and validation of an external additional trial. *Alzheimers Res. Ther.* **15**, 127. https://doi.org/10.1186/s13195-023-01268-9 (2023).
21. Jiang, H. & Nachum, O. In *International Conference on Artificial Intelligence and Statistics* 702–712 (PMLR).
22. Kwon, R., Lee, H., Kim, M. S., Lee, J. & Yon, D. K. Machine learning-based prediction of suicidality in adolescents during the COVID-19 pandemic (2020–2021): Derivation and validation in two independent nationwide cohorts. *Asian J. Psychiatr.* **88**, 103704. https://doi.org/10.1016/j.ajp.2023.103704 (2023).
23. Mathanker, S., Weckler, P., Bowser, T., Wang, N. & Maness, N. AdaBoost classifiers for Pecan defect classification. *Comput. Electron. Agric.* **77**, 60–68 (2011).
24. Chen, T., & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA* 785 (2016).
25. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232. http://www.jstor.org/stable/2699986 (2001).
26. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
27. Lee, S. W. Regression analysis for continuous independent variables in medical research: Statistical standard and guideline of life cycle committee. *Life Cycle*. **2**, e3. https://doi.org/10.54724/lc.2022.e3 (2022).
28. Hwang, S. H. et al. Machine learning-based prediction for incident hypertension based on regular health checkup data: Derivation and validation in 2 independent nationwide cohorts in South Korea and Japan. *J. Med. Internet Res.* **26**, e52794. https://doi.org/10.2196/52794 (2024).
29. Barouch, D. H. Covid-19 vaccines—Immunity, variants, boosters. *N. Engl. J. Med.* **387**, 1011–1020. https://doi.org/10.1056/NEJMra2206573 (2022).
30. Andrew, P. B. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* **30**, 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2 (1997).
31. Huang, L. et al. 1-Year outcomes in hospital survivors with COVID-19: A longitudinal cohort study. *Lancet* **398**, 747–758. https://doi.org/10.1016/s0140-6736(21)01755-4 (2021).
32. Gong, L. et al. Human papillomavirus vaccine-associated premature ovarian insufficiency and related adverse events: Data mining of vaccine adverse event reporting system. *Sci. Rep.* **10**, 10762. https://doi.org/10.1038/s41598-020-67668-1 (2020).
33. Wong, K. K. et al. Menstrual irregularities and vaginal bleeding after COVID-19 vaccination reported to v-safe active surveillance, USA in December, 2020–January, 2022: An observational cohort study. *Lancet Digit. Health* **4**, e667–e675. https://doi.org/10.1016/S2589-7500(22)00125-X (2022).
34. Chapin-Bardales, J., Gee, J. & Myers, T. Reactogenicity following receipt of mRNA-based COVID-19 vaccines. *JAMA* **325**, 2201–2202. https://doi.org/10.1001/jama.2021.5374 (2021).
35. Lee, S. et al. Global and regional burden of vaccine-induced thrombotic thrombocytopenia, 1969–2023: Comprehensive findings with critical analysis of the international pharmacovigilance database. *Eur. J. Haematol.* **113**, 426–440. https://doi.org/10.1111/ejh.14250 (2024).

36. Lee, K. M. N. et al. Investigating trends in those who experience menstrual bleeding changes after SARS-CoV-2 vaccination. *Sci. Adv.* **8**, eabm7201. https://doi.org/10.1126/sciadv.abm7201 (2022).
37. Hwang, J. et al. Predictors of mortality in thrombotic thrombocytopenia after adenoviral COVID-19 vaccination: The FAPIC score. *Eur. Heart J.* **42**, 4053–4063. https://doi.org/10.1093/eurheartj/ehab592 (2021).
38. Hajjo, R., Momani, E., Sabbah, D. A., Baker, N. & Tropsha, A. Identifying a causal link between prolactin signaling pathways and COVID-19 vaccine-induced menstrual changes. *npj Vaccines* **8**, 129. https://doi.org/10.1038/s41541-023-00719-6 (2023).
39. Arora, A. et al. The value of standards for health datasets in artificial intelligence-based applications. *Nat. Med.* **29**, 2929–2938. https://doi.org/10.1038/s41591-023-02608-w (2023).
40. Fernández, A., Garcia, S., Herrera, F. & Chawla, N. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **61**, 863–905. https://doi.org/10.1613/jair.1.11192 (2018).
41. Khan, A. A., Chaudhari, O. & Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation (2023). arXiv:2304.02858.
42. Pottegård, A. et al. Arterial events, venous thromboembolism, thrombocytopenia, and bleeding after vaccination with Oxford-AstraZeneca ChAdOx1-S in Denmark and Norway: Population based cohort study. *BMJ* **373**, 1114. https://doi.org/10.1136/bmj.n1114 (2021).
43. Blix, K. et al. Unexpected vaginal bleeding and COVID-19 vaccination in nonmenstruating women. *Sci. Adv.* **9**, eadg1391. https://doi.org/10.1126/sciadv.adg1391 (2023).
44. Wang, G. et al. Determinants of COVID-19 vaccination status and hesitancy among older adults in China. *Nat. Med.* **29**, 623–631. https://doi.org/10.1038/s41591-023-02241-7 (2023).
45. Granese, R. et al. Effects of SARS-CoV-2 vaccination on menstrual cycle: An Italian survey-based study. *J. Clin. Med.* **12**, 7699. https://doi.org/10.3390/jcm12247699 (2023).

Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00438239) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00509257, Global AI Frontier Lab). Hayeon Lee was supported by grant from the health fellowship foundation (2023). The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Author contributions

Dr D.K.Y. had full access to all data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors have approved the final version of the manuscript before submission. Study concept and design: Y.C., J.P., H.L., H.K., D.K.Y., and J.L.; Acquisition, analysis, or interpretation of data: Y.C., H.L., H.K., D.K.Y., and J.L.; Drafting of the manuscript: Y.C., H.L., H.K., D.K.Y., and J.L.; Critical revision of the manuscript for important intellectual content: Y.C., H.L., H.K., Y.J.L., Y.L., Y.S.C., S.G.Y., J.K., M.R., D.K.Y., and J.L.; Statistical analysis: Y.C., H.L., H.K., D.K.Y., and J.L.; Study supervision: Y.C., H.L., H.K., D.K.Y., and J.L. D.K.Y. is the guarantor for this study. Y.C., J.P., H.K., and Y.J.L. contributed equally as first authors. H.L., D.K.Y., and J.L. contributed equally as corresponding authors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Declarations

Competing interests

The authors declare no competing interests.

Informed consent

This study used deidentified national data, informed consent was waived by KDCA and Kyung Hee University.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91882-4.

Correspondence and requests for materials should be addressed to D.K.Y. or J.L.