Development of Predictive Models for Personalized Warfarin Dose Optimization using Statistical and Machine Learning Approaches

Keetapat Tananchai | Thita Maraya | Tanakorn Yimyorng | Poramad Boontha | Kanjanapron Matapan | Advisor Asst. Prof. Dr. Phimphaka Taninpong

Faculty of Science, Data Science Program, Chiang Mai University

FACULTY OF SCIENCE
CHIANG MAI UNIVERSITY

ABSTRACT

This study aimed to develop and evaluate statistical and machine learning models for individualized warfarin dose prediction using data from the International Warfarin Pharmacogenetics Consortium (IWPC) and hospital-based electronic medical records (EMR). This retrospective study assessed 39 models, including Linear Regression, Random Forest, and Gradient Boosting Regression Trees (GBRT), along with several ensemble approaches. Predictor variables included demographic and clinical characteristics such as ICD-10 diagnostic codes, laboratory test results, and medication transaction records. However, pharmacogenetic factors, including CYP2C9 and VKORC1 genotypes, were available only in the IWPC dataset. Model performance was evaluated using R^2, mean absolute error (MAE), root mean square error (RMSE), and the percentage of predictions within 20% of the actual dose (PW20). In the hospital EMR dataset, Lasso regression achieved the best performance (R^2 = 0.834, MAE = 4.42 mg/week, PW20 = 52.38%). In the IWPC cohort, a stacked ensemble model (RF + SVR + Ridge) showed the best performance in the Asian subgroup (R^2 = 0.262, PW20 ≈ 41%), while the same ensemble approach demonstrated superior performance in the non-Asian subgroup (R^2 = 0.211, PW20 = 38.10%). These findings highlight the importance of predictive modeling for individualized warfarin dose optimization, as the accuracy of dose prediction varied across datasets and patient populations.

INTRODUCTION

Warfarin is a widely used oral anticoagulant in the vitamin K antagonist class and remains important for preventing and treating thromboembolic disorders such as atrial fibrillation, venous thromboembolism, and prosthetic heart valve replacement. Despite the growing use of direct oral anticoagulants, warfarin continues to play a key role in many clinical settings. However, it has a narrow therapeutic index and considerable interindividual dose variability (Thipsumon Raksintham, 2023). Inappropriate dosing may cause serious bleeding or thromboembolic events, making accurate dose prediction essential. Warfarin response is influenced by clinical factors (e.g., age, body weight, medications) and genetic polymorphisms in CYP2C9 and VKORC1, which significantly contribute to dose variability. Therefore, machine learning–based predictive models offer a promising approach to integrate these factors and improve individualized warfarin dose estimation (Nguyen et al., 2021).



OBJECTIVES

This study aims to develop personalized warfarin dose prediction models using secondary data from a public dataset and hospital electronic medical records, and to compare the predictive performance of these models across different dataset

METHODOLOGIES

Datasets This study used two datasets:

1. IWPC Dataset : The IWPC dataset is a publicly available dataset from the International Warfarin Pharmacogenetics Consortium, including 5,503 samples (IWPC Consortium, 2009).
2. Hospital Dataset : Clinical data from 100 patients obtained from hospital electronic medical records.

The IWPC dataset was divided into training (70%) and test (30%) sets, while the hospital dataset was divided into training (80%) and test (20%) sets for model evaluation.

IWPC	Training Set	Test Set	Total Samples	Proportion
Asian	1,294	324	1,618	29.40%
Non-Asian	3,108	777	3,885	70.60%
Total	4,402	1,101	5,503	100.00%

Table 1 : IWPC dataset : Experiment Setup

Hospital	Training	Test Set	Total	Proportion
Patients	80	20	100	100.00%

Table 2 : hospital dataset

- Features include demographics, clinical factors, laboratory values, and drug interactions.
- Only the IWPC dataset contains only pharmacogenetic factors (CYP2C9 and VKORC1) related to warfarin.
- The hospital EMR dataset reflects real-world clinical practice.

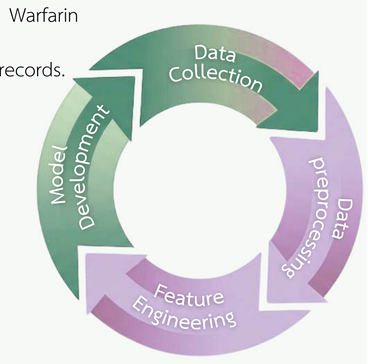


Figure1 : Workflow of the Machine Learning Methodology for Warfarin Dose Prediction

DATA PREPROCESSING

Data Integration

Demo graphic + Drug + diagnosis + Lab

Drug Extraction & Classification

Drug → icode, strength, dosage instructions, visit date → Warfarin and Statin

Prescription records were aggregated by patient and visit date to identify medications used at each visit.

ICD-10 Diagnosis Processing

Extract the first 3 digits of ICD-10 codes for disease grouping and mapping.

ex. I48.9 → I48 → Atrial Fibrillation (AF)

Drug Dose Calculation

- Normalize prescription text (trim spaces and convert to lowercase)
- Extract drug strength, tablet amount and dosing days
- Compute Statin daily dose (mg/day)

  Dose = Strength × Tablets × Frequency per day
- Compute Warfarin weekly dose (mg/week)

  Dose = Strength × Tablets × Days per week

Examples :

Drug	Prescription	Strength	Tablets	Days per Week	Calculation	Dose
Statin	1X1HS	20 mg	1	1	20 × 1 × 1	**20 mg/day**
Warfarin	1x1HS จ-ส	5 mg	1	6	5 × 1 × 6	**30 mg/week**

Visit Timeline Construction

- Identify five visits from drug and INR data
- Visit 1: The first time the patient receives both Warfarin and Statin.
- (visit1 -visit5) This study used the data in visit4 which inr is in target.

FEATURE ENGINEERING

Derived Features

M / W → BMI² : BMI squared; BSA : Body Surface Area; Age midpoint

Drug Interaction Feature Construction

Identify interacting medication → Binary encoding → DrugInteraction count

Medication	Value
drug_other_1	0
Statin	1
drug_other_2	0

0+1+0=1

Converted categorical variables → one-hot encoding

Gender	Smoking	Drinking
Male	Yes	No
Female	No	Yes

Gender_Male	Gender_Female	Smoking_Yes	Smoking_No	Drinking_Yes	Drinking_No
1	0	1	0	0	1
0	1	0	1	1	0

Handled missing values → Missing indicator method

Identify missing values → Create a binary indicator variable → Create a missing indicator variable

Lab
35
NaN
28

lab1	lab1_missing
1	0
0	1
1	0

Feature selection

- **Correlation analysis** to identify informative predictors.
- **Multicollinearity** check to remove redundant variables.
- **Regularization methods** (Lasso, ElasticNet) to shrink less important features.
- **Tree-based** feature importance to retain the most predictive variables.

Machine learning models

Models used in this study were categorized based on their learning algorithms, including Linear, SVM, PLS, Decision Tree, Ensemble, KNN, ANN, and Probabilistic models. These models apply different learning mechanisms and were compared to evaluate their predictive performance. In total, 39 models were analyzed using the study datasets.

MACHINE LEARNING MODELS	
Linear Models	Support Vector Machines (SVM)
Partial Least Squares	Decision Tree Models
Ensemble Learning Models	K-Nearest Neighbors (KNN)
Artificial Neural Networks	Probabilistic Models

EVALUATION METRICS

MAE → Mean absolute prediction error :

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

R^2 → Proportion of variance explained by the model :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

RMSE → RMSE – Root mean squared error (penalizes large errors) :

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

PW20% → Percentage of patients whose predicted dose was within ±20% of the actual stable dose.

$$PW20 = \frac{1}{N}\sum_{i=1}^{N} I\left(\frac{|\hat{y}_i - y_i|}{y_i} \le 0.20\right) \times 100$$

RESULTS

Model Performance for Warfarin Dose Prediction Using IWPC Dataset

	Asian					Non-Asian				
	Model	MAE (mg/week)	RMSE (mg/week)	R^2	PW20 (%)	Model	MAE (mg/week)	RMSE (mg/week)	R^2	PW20 (%)
Best Single Model	XGBRegressor_100_Lighter	6.671	9.252	0.214	39.81	GBRT_AbsoluteLoss	10.836	14.191	0.197	38.1
Best Stacking Ensemble Model	RF, SVR (base) and Ridge (meta)	6.602	8.967	0.262	40.74	RF, SVR (base) and Ridge (meta)	10.922	14.066	0.211	38.1

Table 3: The Best Machine Learning Model Performance for Warfarin Dose Prediction Using the IWPC Dataset

The comparison of predictive models showed that the Stacking Ensemble model (RF, SVR, and Ridge) achieved the best performance in the Asian group, with R^2 = 0.262 and RMSE = 8.967 mg/week, slightly outperforming the single model (XGBRegressor). In the Non-Asian group, the GBRT (AbsoluteLoss) model provided the best result with R^2 = 0.197.

The difference in predictive performance between Asian and Non-Asian populations may be explained by variations in the frequency of genetic variants related to Warfarin response, particularly VKORC1 and CYP2C9. These variants occur at different rates across ethnic groups. In particular, Asian populations often carry VKORC1 variants associated with higher sensitivity to Warfarin, which may help machine learning models better capture the relationship between genetic factors and optimal dosage.

Model Performance for Warfarin Dose Prediction Using Hospital EMR Data



Figure 2. Performance Comparison of the Top 5 Models for Warfarin Dose Prediction Using the Hospital EMR Dataset

The model comparison indicates that regression-based models achieved the best overall predictive performance. Linear Regression produced the highest R^2 value (0.834), indicating the strongest ability to explain the variability in warfarin dose. The RMSE and MAE values were very similar across models, suggesting comparable prediction errors. Therefore, PW20 was also considered to assess the proportion of predicted doses within ±20% of the actual dose. Although Ridge Regression (α = 1) achieved the highest PW20, Linear Regression was selected as the final model because it demonstrated the highest explanatory power (R^2) while maintaining comparable error metrics, indicating a better overall balance between accuracy and interpretability.

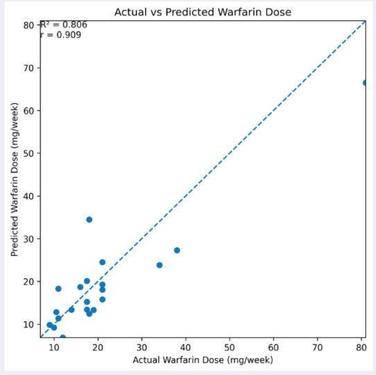


Figure 3. Actual vs. predicted warfarin dose using the Lasso regression model.

Lasso regression was selected as the best-performing model based on its predictive performance. The model produced the following simplified prediction equation:

Dose = (1.636817 + (-0.064853 × smoking_type_id) + (0.044081 × drinking_type_id) +.....+ (0.101916 × Total_bilirubin_missing) + (0.359569 × Sex_binary_missing)) - 1

Differences in predictive performance between the IWPC and hospital EMR datasets may be related to variations in ethnicity, availability of pharmacogenetic variables, and dataset characteristics, the hospital dataset consists only of Thai patients and does not include genetic information. Despite the smaller dataset and absence of genetic variables, the hospital-based model achieved higher predictive performance, suggesting that models developed from local clinical data may better capture population-specific characteristics and treatment patterns.

CONCLUSION

This study developed and evaluated predictive models for individualized warfarin dose estimation using two different data sources. In the hospital EMR dataset, Lasso regression was identified as the most suitable model for dose prediction. In contrast, a stacked ensemble model combining Random Forest, Support Vector Regression, and Ridge regression demonstrated the best performance for the IWPC dataset. These findings suggest that the optimal predictive model may vary depending on the characteristics of the dataset and the population being studied.

REFERENCES

Raksintham, T. (2023). Factors associated with over-therapeutic range INR of warfarin-treated patients in warfarin clinic, Prachuap Khiri Khan Hospital. Journal of Nakornping Hospital, 16(1).

Nguyen, T., Zhang, Y., Nguyen, H., Nguyen, D., & Nguyen, T. (2021). Comparison of multivariate linear regression and machine learning algorithm developed for prediction of precision warfarin dosing in a Korean population. Journal of Clinical Pharmacy and Therapeutics, 46(3), 673–685.

International Warfarin Pharmacogenetics Consortium. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. New England Journal of Medicine, 360(8), 753–764.