Check for updates

Original Article

Machine Learning Models for Warfarin Dose Prediction: A Systematic Review of Performance and Clinical Utility

Journal of Applied Pharmaceutical Science
16(9) 941–952, 2026
© The Author(s) 2026
Article reuse guidelines:
in.sagepub.com/journals-permissions-india
DOI: 10.1177/22133354261466853
journals.sagepub.com/home/phs
Sage

Rini Noviyani[1], Han N. Nguyen[2], Khang L.T. Duong[3], Trinh V. Ngo[2], Nhi T.N. Ngo[4], I Gusti Ngurah Lanang Wijayakusuma[5], Ida Ayu Alit Widhiartini[6] and Khanh N.C. Duong[2]

Abstract

Warfarin therapy is complicated by its narrow therapeutic index and significant inter-individual variability. This study aimed to systematically review the performance of machine learning (ML) for predicting warfarin doses in clinical settings, compared with conventional methods, and to assess the quality of evidence. We searched PubMed, Embase, and Web of Science until December, 2024, and included studies developing ML models for warfarin dosing in clinical settings. The performance metrics of ML models were evaluated using predictive accuracy, model fit, and clinical utility. Risks of bias were assessed using the Prediction Model Risk of Bias Assessment Tool. Thirty-five studies were included. In terms of ML models, the majority of included studies evaluated artificial neural networks (n = 24), followed by ensemble models (n = 19), and support vector machines (n = 18). Regarding performance, ML models demonstrated superiority across all three metrics: predictive accuracy (16 of 18 studies), model fit (5 of 7 studies), and INR-related clinical utility (all 4 studies). A high risk of bias was identified in 94% of studies, mainly due to analysis-domain limitations, including missing-data handling, insufficient events per predictor, unclear predictor selection, and limited external validation. In conclusion, ML-based approaches may improve warfarin dose prediction compared with conventional methods, but the current evidence should be interpreted cautiously. Future studies should prioritize transparent reporting, robust model development, external validation, and clinically relevant outcomes before these models can be recommended for routine clinical implementation.

Keywords

Machine learning, warfarin, anticoagulants, systematic review, dose prediction

Received 19 March 2026; **accepted** 08 June 2026

Introduction

Warfarin is a widely prescribed oral anticoagulant, used primarily for the prevention and treatment of thromboembolic disorders.[1,2] Despite its efficacy, warfarin therapy is complicated by its narrow therapeutic window and significant inter-individual variability. This variability is influenced by factors such as age, weight, diet, comorbidities, and genetic polymorphisms. Incorrect dosing can lead to severe adverse clinical outcomes, including major bleeding or thrombotic events, making accurate dose prediction critical for patient safety.[3]

Historically, clinicians have relied on clinical and pharmacogenetic algorithms, such as those developed by Gage et al. and the International Warfarin Pharmacogenetics Consortium (IWPC), to determine initial warfarin dosing.[4,5] Although these models improve upon empirical methods, their predictive accuracy and applicability are often limited,

Rini Noviyani and Han N. Nguyen contributed equally to this work.
[1] Department of Pharmacy, Faculty of Mathematics and Natural Sciences, Udayana University, Denpasar, Bali, Indonesia
[2] Department of Economic, Social, and Administrative Pharmacy, Faculty of Pharmacy, University of Health Sciences, Vietnam National University Ho Chi Minh City, Vietnam
[3] Department of Computer Science, University of Information Technology, Vietnam National University Ho Chi Minh City, Vietnam
[4] Mahidol University Health Technology Assessment (MUHTA) Program, Mahidol University, Bangkok, Thailand
[5] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Udayana University, Denpasar, Bali, Indonesia
[6] Department of Pharmacology and Therapy, Faculty of Medicine, Udayana University, Denpasar, Bali, Indonesia

Corresponding author:
Khanh N.C. Duong, Department of Economic, Social, and Administrative Pharmacy, Faculty of Pharmacy, University of Health Sciences, Vietnam National University Ho Chi Minh City, YA1 Administrative Building, Hai Thuong Lan Ong Street, Dong Hoa Ward, Ho Chi Minh City 700000, Vietnam.
E-mail: dnckhanh@uhsvnu.edu.vn

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-Commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the Sage and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

particularly in diverse populations. This highlights the importance of developing more precise and personalized predictive tools.

In recent years, machine learning (ML), a subfield of artificial intelligence, has emerged as a powerful approach for modeling complex, nonlinear interactions among numerous predictors.[6,7] There are several ML models commonly studied, including artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and other ensemble models (EMs).[6–8] ANNs are structured into input, hidden, and output layers, with weighted connections between nodes that integrate raw features to generate a predicted outcome.[8,9] DTs classify data through recursive partitioning, in which each node sequentially splits observations based on specific feature values, whereas random forests (RFs) aggregate predictions from multiple DTs via majority voting or averaging to improve accuracy and reduce overfitting.[8–10] SVMs operate by projecting data into a higher-dimensional space via a kernel function, then identifying the optimal hyperplane that maximizes the margin between classes, conferring robustness even when the number of features exceeds the number of training samples.[8,11,12]

The use of ML to predict warfarin dosing is steadily increasing. A systematic review, with data updated through March 2022, on the application of ML to individualized dosing found that 23 studies used ML for warfarin dosing.[7] Although several studies and reviews explored ML for warfarin dosing, they often have limitations. For instance, a systematic review by Zhang et al. (2022)[7] focused on the technical aspects of model development but did not comprehensively evaluate the comparative performance of ML against conventional algorithms. Two reviews by Iancu et al. (2023)[13] and Li et al. (2024)[6] did not restrict the drug type to warfarin, resulting in the inclusion of studies on various medications that may limit the applicability of their findings to warfarin dosing prediction. In addition, neither review provided sufficient performance metrics to enable meaningful comparison of the evaluated predictive models.[6,7,13] Furthermore, existing reviews have given limited attention to the clinical readiness of these models, either by analyzing external validation results, reporting clinical outcomes, or assessing the applicability of the ML models using the Prediction Model Risk of Bias Assessment Tool (PROBAST).[6,7,13]

To address these gaps, we conducted a systematic review of studies that employed ML for warfarin dose prediction in the clinical setting. This study aimed to summarize the characteristics of the ML models used, their predictors, and the type of performance metrics reported in the literature. Furthermore, we aimed to compare the performance of ML-based approaches against conventional methods (e.g., empirical dosing or non-ML regression algorithms). Our findings provide a comprehensive overview of the current landscape, highlighting key evidence to guide future development and clinical implementation.

Materials and Methods

We conducted this systematic review and reported our findings in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[14] The review protocol was registered with the Open Science Framework (registration ID: osf.io/fg3dn). No major deviations from the original protocol occurred.

Search Strategy and Selection Criteria

We performed a systematic search of three online databases: PubMed, Embase, and Web of Science, for records up to December 24, 2024. The search strategy was designed by combining the keywords "machine learning" and "warfarin" with their respective synonyms. The initial search strategy was developed for one database and then adapted for other databases. There were no restrictions on publication date or language (see Supplemental Material Table S1–S3 for full search strategies).

We screened the identified records in two phases: first by title and abstract, then by full text. At each phase, four reviewers (R.N., H.N., K.L.T.D., T.N.) worked in two independent pairs to assess the records using the Rayyan platform (https://www.rayyan.ai/). The studies were included in this systematic review if they (a) developed, tested, or validated ML models in warfarin dosing and (b) reported performance metrics for the ML models. Studies were excluded if they (a) were pharmacokinetic/pharmacodynamic (PKPD) studies; (b) were published as letters, editorials, or conference abstracts; and (c) were not published in English. Disagreements during the screening process were resolved through discussion among the reviewers or by consulting a fifth reviewer (K.N.C.D.).

Data Extraction

Four reviewers (R.N., H.N., K.L.T.D., T.N.), working in two independent pairs, extracted data from all eligible studies using a standardized Microsoft Excel template. Each study was extracted independently by two reviewers, and disagreements were resolved through discussion or consultation with a fifth reviewer (K.N.C.D.). We extracted the following data: (a) study characteristics: first author, year of publication, country, study design, data source, comparison type, outcome measured, number of patients, population characteristics, warfarin indications, and patient race/ethnicity; (b) ML model characteristics: ML algorithms used, number and type of predictors, methods for predictor selection, model performance metrics, approaches for handling missing data, and use of internal or external validation; and (c) comparative performance data: the specific models being compared (e.g., ML vs. conventional), their respective performance metric values, and the study's conclusions on which model performed better.

Outcome Measures

The primary outcomes of this review were the performance metrics of the ML methods as reported in the included studies. We categorized these into three main types: predictive accuracy, model fit, and clinical utility. Predictive accuracy metrics measure the error between predicted and actual doses, including mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). Model fit was typically assessed using the coefficient of determination (R^2). Clinical utility refers to numerical evaluations of the potential real-world impact of the models, specifically their direct effects on improving therapeutic efficacy and patient safety. Examples of such metrics include time to target International Normalized Ratio (INR), time in the therapeutic range, and the incidence of bleeding or thrombotic events.

Data Synthesis

We synthesized the findings narratively. Descriptive statistics and visualizations, such as tables and plots, were used to summarize the extracted data. Microsoft Excel and R were used to generate illustrative charts as needed. We did not perform meta-analysis due to high heterogeneity across studies (e.g., diverse ML algorithms, different predictors, variable outcome definitions, and inconsistent performance metrics).

Risk of Bias Assessment

We assessed the risk of bias of included studies using the PROBAST.[15] The tool uses a 20-item checklist to evaluate the risk of bias and applicability of ML models across four domains: participants, predictors, outcome, and analysis. Each item was answered with "Yes," "Probably yes," "No," "Probably no," or "No information." For each domain, if the answer was "No" or "Probably no" for at least one question, the domain was evaluated to have "high risk of bias." For each study, if at least one of its domains had a high risk of bias, the overall risk of bias assessment for that study was also "high." We adapted the checklist for ML-based studies by removing irrelevant questions and modifying others based on previously published systematic reviews[6,7] and guidance on prediction model development[16] (see Supplemental Material Table S4). We also evaluated the applicability of the models across three domains: participants, predictors, and outcomes. Based on these assessments, each study was assigned an overall risk of bias and overall applicability rating of "low," "high," or "unclear."[15] Two independent reviewers (H.N., N.N.) conducted the assessment, and a third reviewer (K.N.C.D.) resolved any conflicts.

Results

Study Selection

Our search across PubMed, Embase and WoS retrieved 12,091 articles. After removing 4,088 duplicates, we screened the titles and abstracts of the remaining 8,003 articles, resulting in the retrieval of 127 studies for full-text screening. Ninety-two full texts were excluded for the following reasons: wrong outcome (50 studies), wrong intervention (24 studies), ineligible publication type (letters/abstracts/editorials, 8 studies), not published in English (7 studies), and wrong population (3 studies). Finally, 35 articles met the inclusion criteria and were included in this systematic review.[17–51] The complete study selection process is detailed in the PRISMA flowchart (see Figure 1).

Characteristics of Included Studies

The 35 included studies were published between 2004 and 2024, with a notable increase in recent years; 18 studies (51%) were published within the past 5 years.[19,34–50.] Most studies were multinational (14 studies, 40%),[17,18,24–27,30,37–39,43,46–48] with a significant number originating from China (13 studies, 37%).[18,25,29,32–35,38,42,44,49–51] All studies used retrospective cohort designs. Fourteen studies (40%) used local hospital or clinic databases,[20–23,31,32,34–36,40,41,44,45,49] 12 (34%) used public databases,[17,24,26,27,29,30,33,42,43,46,50,51] 8 (23%) used mixed databases[18,19,25,37–39,47,48] and one did not describe the database used.[28] The IWPC database was the most frequently used public resource (13 studies, 37%).[17,18,24–27,30,37–39,43,46,47] The median number of participants was 1,501 (ranging from 148[20] to 28,232[48]), and the populations were predominantly Asian (16 studies, 46%).[18,22,25,28,29,32–35,38,40,42,44,49–51] The primary outcomes measured were weekly (16 studies, 46%)[17,23–28,30,31,37,39–41,46–48] and daily maintenance dose (12 studies, 34%).[18,20,22,29,32,33,35,38,42,49–51] Most studies (27 studies, 77%) did not specify the indications for warfarin therapy.[17–28,30–32,35–40,43,45–49] The warfarin indications for the remaining studies included heart valve replacement (seven studies, 20%)[29,33,34,42,44,50,51] and multiple indications (venous thromboembolism, atrial fibrillation, or valvular heart disease) (one study, 3%).[41]

Further details are summarized in Table 1 and provided in Supplemental Material Table S5.

Characteristics of ML Models

A wide range of ML algorithms was employed. ANNs were the most common (24 studies, 69%),[17–20,22–25,27–33,38,39,41–44,46,50,51] followed by EMs (19 studies, 54%)[17–19,21,22,25,30–32,35–39,41,43–45,49] and SVMs (18 studies, 51%).[17,18,21,22,25,26,30–32,35,37–41,43,44,49] The number of predictors used to train the models ranged from 3[20] to 44,[45] with a median of 10. These predictors included patient demographic, clinical, and warfarin-related factors. Particularly, patient demographic factors included age (32 studies, 91%),[17–22,24–33,35–49,51] gender (21 studies, 60%),[18,19,22,23,25,26,28,30,32,35–38,40,41,43–48] and race/ethnicity (10 studies, 29%).[17,24,26,30,31,37,39,43,46,48] Clinical factors included concomitant medications (25 studies, 71%),[17–21,23–27,30,32,34–39,41,43,45–48,51] patient weight (24 studies, 69%),[17–19,21–25,27,29–32,34–39,41–43,46,48]

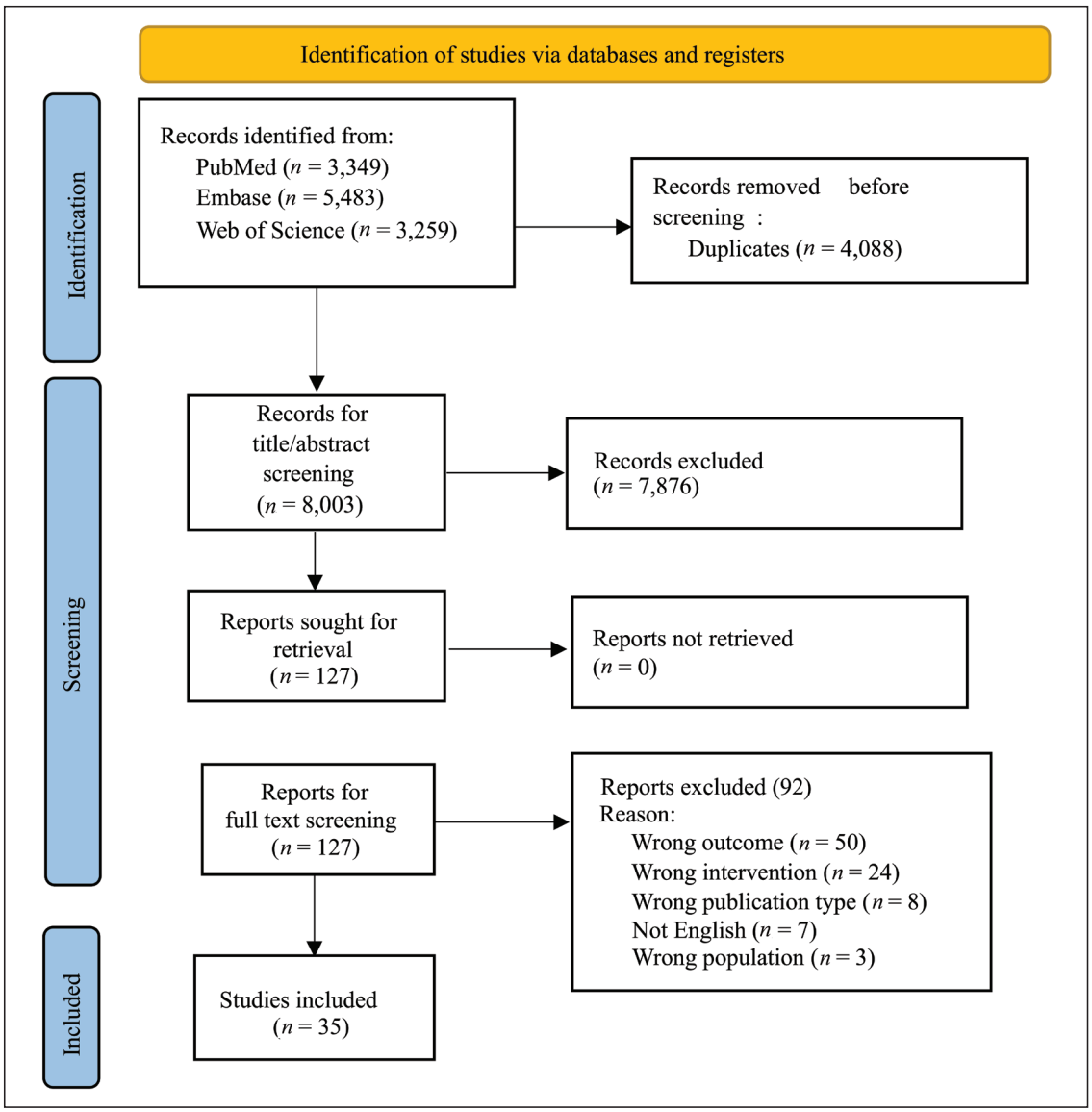

Figure 1. PRISMA Flowchart of Study Selection Process.

Table 1. General Characteristics of Included Studies.

Characteristic	Number of Studies	Percentage
Total Included Studies	35	100
Year of Publication	35	100
2004–2010	1	3
2011–2015	7	20
2016–2020	9	26
2021–2024	18	51
Country		
China	13	37
India	2	6
South Africa	2	6
South Korea	2	6
United States	2	6
Multi-country	14	40
Others[a]	7	20

(Table 1 continued)

(Table 1 continued)

Characteristic	Number of Studies	Percentage
Study Design		
Retrospective cohort study	35	100
Database		
Hospital/Clinic/Medical center database	14	40
Public database	12	34
Mixed databases[b]	8	23
Not reported	1	3
Comparison Type		
Machine learning versus conventional methods[c]	23	66
Machine learning versus machine learning	22	63
Machine learning alone	4	11
Outcome Measured		
Weekly maintenance dose	16	46
Daily maintenance dose	12	34
Daily dose at discharge	2	6
Others[d]	5	14
Number of Patients (Median (Min-Max)	1,501 (148–28,232)	
Population		
Adults	15	43
Unspecified	20	57
Warfarin Indications		
Heart valve replacement	7	20
Mixed thromboembolic indications	1	3
Unspecified	27	77
Race/Ethnicity		
Asian	16	46
African	3	9
Hispanic/Latin	2	6
White	1	3
Unspecified	12	34

Notes: Percentages may not sum to 100 as studies could report multiple characteristics (e.g., comparison types).
[a]The "Others" category under Country includes countries represented by only a single study.
[b]Represents studies that utilized a combination of public and hospital/clinic databases.
[c]Conventional methods include clinician-estimated doses and non-machine learning regression algorithms.
[d]The "Others" category under Outcome measured includes outcomes that were each reported in only a single study.

genetic factors (22 studies, 63%),[17,18,21,23–25,27,28,30–32,34–39,43,44,46,47,49] and comorbidities (20 studies, 57%).[19,21,22,25–28,30,31,36,37,41–46,48–50] The most frequently included interacting co-medication was amiodarone (22 studies, 63%).[17–21,23,24,26,27,30,32,34,36–39,41,43,45–48] Among genetic factors, variants in vitamin K epoxide reductase complex subunit 1 (VKORC1) (22 studies, 63%)[17,18,21,23–25,27,28,30–32,34–39,43,44,46,47,49] and cytochrome P450 family two subfamily C member 9 (CYP2C9) (21 studies, 60%)[17,18,21,23–25,27,28,30–32,35–39,43,44,46,47,49] were the most frequent. Warfarin-related factors, such as indication for use (16 studies, 46%)[23,25,26,30,31,35–40,43–46,49] and the initial or current dose (7 studies, 20%),[20,34,40,42,45,48,50] were also used as predictors (see Table 2 for the summary of predictors and Supplemental Material Tables S6–S7 for details of the predictors in each study).

Model Development and Validation

The methods for predictor selection varied across the studies. For the initial selection of candidate predictors, 16 studies (46%) selected them based on prior knowledge or the literature,[18–23,29–31,38,39,41,42,45,48,51] while 14 studies (40%) did not report their selection method.[24,27,28,32,34,36,37,40,43,44,46,47,49,50] To determine the final set of predictors, 13 studies (37%)[17,22,25,26,30,31,39–41,44–46,48] included all candidate predictors into the final model, seven studies (20%),[20,21,28,33,42,50,51] five studies (14%),[24,27,34,35,43] and five studies (14%)[18,24,32,38,43] used univariable analysis, stepwise selection, and measuring model performance to select final predictors, respectively. The primary metrics used to evaluate model performance were the MAE (30 studies, 86%),[17–19,22–39,41–43,45–47,49–51] the percentage of predicted dose in a predefined range (24 studies, 69%),[17,18,23–25,27,29–33,35–39,41,42,45–47,49–51] and the coefficient of determination or correlation coefficient (16 studies, 46%).[18,21,23,24,27,28,32,34–36,38,40,43,45,49,51] Over half of the studies (18 studies, 51%)[18,20,22,23,27–29,32–34,38,40,44,45,47,49–51] did not report how they handled missing data. For those that did, the most common methods were complete case exclusion (7 studies, 20%)[17,24,25,31,42,46,48] and multiple imputation (7 studies, 20%).[26,30,35,37,39,41,43] While most studies performed internal

validation (31 studies, 89%),[17–30,32–36,38–46,49–51] only 11 (31%) conducted external validation to test model generalizability on a separate dataset.[19,25,29,33,34,40–42,48,50,51] Table 2 summaries the performance metrics of the included studies, and Supplemental Materials Table S8 provides the details of the performance metrics used in each included study.

Performance of ML Models

Of the 35 included studies, 23 (66%) compared ML models against conventional methods[17–20,22,25–28,30,32–41,43,44,48] and 22 (63%) compared between ML models.[17–19,21,22,25,30–32,35,37–39,41–45,47,49–51] The conventional methods compared include physician prediction and non-ML regression algorithms. The performance was assessed based on predictive accuracy, model fit, and clinical utility (Table 3).

Predictive Accuracy

Predictive accuracy, which measured the error between predicted and actual maintenance doses, was the most common evaluation method. The most frequently used metrics included MAE, MSE, RMSE, and the percentage of predicted doses within a predefined ideal range. The MAE for ML models

Table 2. Characteristics of Machine Learning Modeling Used in Included Studies.

Characteristic	Number of Studies	Percentage
Machine Learning Techniques Used		
Artificial neural networks	24	69
Ensemble model	19	54
Support vector machine	18	51
K-nearest neighbors	6	17
Multivariate adaptive regression splines	6	17
Ridge regression	5	14
LASSO regression	3	9
Others[a]	4	11
Number of Predictors (Median (Min-Max))	10 (3–44)	
Predictors		
Patient Demographic Predictors		
Age	32	91
Gender	21	60
Race/Ethnicity	10	29
Patient Clinical Predictors		
Coadministrations	25	71
Weight	24	69
Genetic factor	22	63
Comorbidities	20	57
Height	17	49
Laboratory test results	16	46
Smoking status	12	34
Body surface area	4	11
Heart failure severity	3	9
Warfarin-related Predictors		
Warfarin indications	16	46
Initial or current warfarin dose	7	20
Target INR	7	20
Starting time of anticoagulants	4	11
Others[b]	12	34
Candidate Predictor Selection		
Prior knowledge/literature	16	46
Stepwise selection	2	6
Predictors with sufficient data	2	6
Univariable analysis	1	3
Not reported	14	40
Final Predictor Selection		
Including all candidate predictors	13	37
Univariable analysis	7	20
Stepwise selection	5	14
Measuring performance	5	14

(Table 2 continued)

(Table 2 continued)

Characteristic	Number of Studies	Percentage
Multivariable analysis	2	6
Recursive feature elimination cross-validation	2	6
Not reported	1	3
Others[c]	4	11
Performance Metrics		
Mean absolute error	30	86
Percentage of ideal dose	24	69
Coefficient of determination/Correlation coefficient	16	46
Mean square error/Root mean squared error	15	43
Measure of under/overdosing	9	26
Others[d]	13	37
Missing Data Handling		
Case exclusion	7	20
Multiple imputation	7	20
Single imputation	3	9
Not reported	18	51
Internal Validation		
Yes	31	89
No	4	11
External Validation		
No	24	69
Yes	11	31

Notes: INR: International normalized ratio. Percentages may not sum to 100 as studies could report multiple characteristics (e.g., machine learning techniques, performance metrics).

[a]The "Others" category under Machine learning techniques used includes techniques that were each reported in only a single study.

[b]The "Others" category under Predictors includes predictors that were each reported in only a single study.

[c]The "Others" category under Final Predictor Selection includes approaches that were each reported in only a single study.

[d]The "Others" category under Performance metrics includes metrics that were each reported in only a single study.

ranged from 0.12 mg/week[32] to 17.59 mg/week.[41] The MSE ranged from 0.08[18] to 11.77 mg/week[35] and the RMSE varied from 0.13[32] to 17.92 mg/week.[43] The percentage of predicted doses within a predefined ideal range varied between 30.0%[32] and 79.1%[50] (see Supplementary Materials Table S9–S10). In the 18 studies that directly compared predictive accuracy, ML models outperformed conventional algorithms across at least one metric in 16 (89%).[17–20,22,25–27,30,32,33,35,38,39,41,43] Specifically, ML models achieved lower MAE, MSE, or RMSE in 15 of 17 comparisons[17–19,22,25–27,30,32,33,35,38,39,41,43] and a higher percentage of ideal dosing in 9 of 13 comparisons.[17,18,30,32,33,35,38,39,41] Despite this strong trend, a conventional MLR model was reported to have higher percentage of ideal dosing than an ML model in four studies[25,27,36,37] (see Table 3 and Supplementary Materials Tables S9–S10 for details of these comparisons).

Model Fit

Model fit is typically reported as the coefficient of determination (R^2), which indicates the proportion of dose variability explained by the model, or as the correlation coefficient (R), which measures the strength of the linear association between actual and predicted doses. The explanatory power of ML algorithms varied significantly, with R^2 values ranging from 0.06[18] to 0.95[40] (see Supplementary Materials Table S11). In the seven studies that made a direct comparison, ML models exhibited a higher R^2 or R value than conventional methods in five of them (71%) (see Table 3).[18,32,35,38,43]

Clinical Utility

Four studies (11%) evaluated the potential clinical impact of the models by assessing patient-centric outcomes.[28,44,45,48] The results consistently favored ML-based approaches in terms of INR-related metrics. Compared with conventional dosing, ML models were found to help patients achieve their target INR more effectively (all three studies)[28,44,45] and to reduce the time to reach the first therapeutic INR (in both studies).[28,44] In terms of adverse events, which were assessed in one study, the ML-based approach was associated with a lower rate of adverse drug reactions; however, this result was not statistically significant ($p = 0.06$)[28] (see Table 3 and Supplementary Materials Table S12).

Quality Assessment

The overall risk of bias across the included studies was high, with 33 studies (94%) judged to be at a high risk of bias.[17–36,38,40–51] The most frequent concerns were in the "Analysis" domain, with 26 studies (74%) rated as high risk[17–28,30–32,35,36,38,42–49] and 7 studies (20%) rated as unclear risk.[29,33,34,39,40,50,51] The main reasons for this included

Table 3. Performance of Machine Learning Models Compared to Conventional Methods.

	Predictive Accuracy			Model Fit		Clinical Utility		
		Specific Metric			Specific Metric		Specific Metric	
Metrics	Total	MAE/MSE/ RMSE	Percentage of Ideal Dose	Total	R^2/R	Total	INR-related Metrics	Rates of AERs
ML compared to conventional methods (no. studies (%))	18 (100)	17 (100)	13 (100)	7 (100)	7 (100)	4 (100)	4 (100)	1 (100)
ML has higher performance than conventional methods (no. studies (%))	16 (89)	15 (88)	9 (69)	5 (71)	5 (71)	4 (100)	4 (100)	0 (0)
ML has lower performance than conventional methods (no. studies (%))	2 (11)	2 (12)	4 (31)	1 (14)	1 (14)	0 (0)	0 (0)	0 (0)
ML is equal performance conventional methods (no. studies (%))	0 (0)	0 (0)	0 (0)	1 (14)	1 (14)	0 (0)	0 (0)	1 (100)

Note: ML has lower performance than conventional methods accounted for all metrics. INR: International Normalized Ratio, MAE: Mean absolute error, ML: Machine learning, MSE: Mean squared error, RMSE: Root mean squared error, AERs: Adverse event reactions, R^2: Coefficient of determination, R: Correlation coefficient.

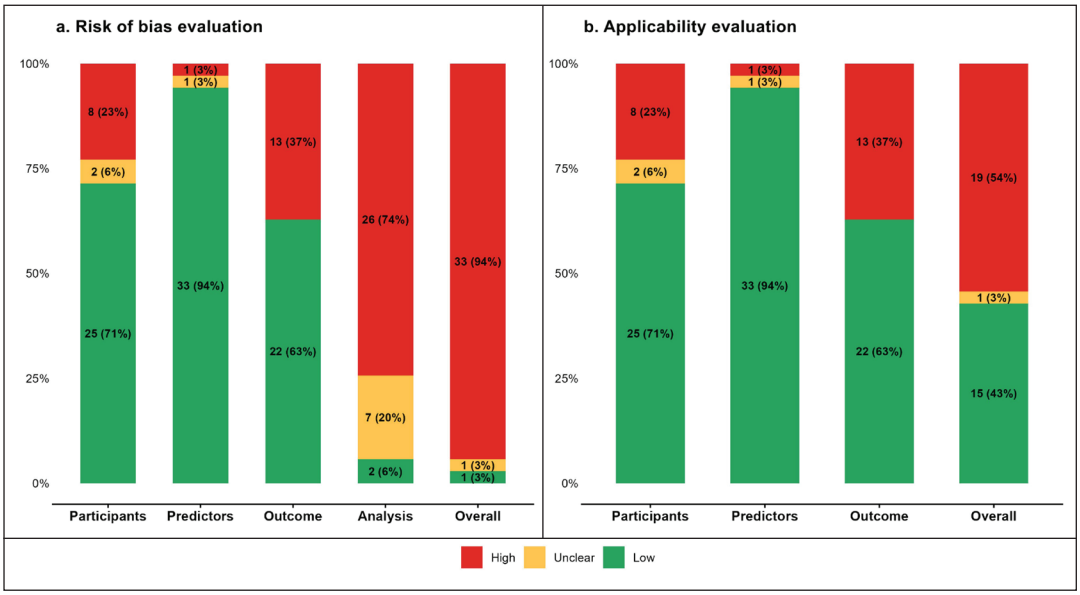

Figure 2. PROBAST Assessment of the Included Studies. (a) Risk of Bias Evaluation for Each Domain and Overall Judgment. (b) Applicability Concerns Evaluation for Each Domain and Overall Judgment. The Numbers Inside the Bars Indicate the Number of Studies, and the Values in Parentheses Represent the Percentage.

inadequate reporting or inappropriate methods for handling missing data (28 studies, 80%),[17–25,27–29,31–34,36,38,40,42,44–51] an insufficient number of outcome events relative to the number of predictors (17 studies, 49%),[18,20–23,26–28,31,32,35,36,38,43,45,47,49] and unclear or inappropriate methods for predictor selection (15 out of 24 development studies—defined as studies with no external validation, 62%).[17,20,21,24,27,28,32,35,36,39,43,44,46,47,49] In contrast, the "Predictors" domain was generally strong, with 33 studies (94%) rated as low risk of bias.[17–19,21–47,49–51] The detailed assessment using the PROBAST tool is summarized in Figure 2 and Supplemental Materials Table S13–S14.

We also evaluated the models' applicability to the review question. Concerns regarding the applicability of the "Predictors," "Participants," and "Outcome" domains were generally low in 94% (33/35),[17–19,21–47,49–51] 71% (25/35),[17–22,24–27,30–32,36–39,41,43–49] and 63% (22/35) of studies,[17,19,21–25,27–32,34,36,37,39,42,47,49–51] respectively. Overall, 15 of the 35 studies (43%) were judged to have low concerns regarding applicability[17,19,21,22,24,25,27,30–32,36,37,39,47,49] (see Figure 2a and 2b).

Discussion

This systematic review synthesized findings from 35 studies that applied ML for warfarin dose prediction. Our review indicated that ML models consistently outperformed conventional methods across performance metrics, including predictive accuracy, model fit, and clinical utility. This demonstrates the substantial potential of ML to enhance dosing precision and improve patient safety.

However, this promising performance is undermined by significant methodological weaknesses in the current literature. A critical finding of this review is that 94% of the included studies had a high risk of bias, with the "Analysis" domain being the primary source of concern. Common flaws included inadequate reporting of methods for handling missing data, non-systematic approaches to predictor selection, and a lack of external validation. These shortcomings compromise the reliability and generalizability of the reported findings, hindering the translation of these models into real-world clinical settings.

Furthermore, only a minority of studies (4 of 35; 11%) assessed the clinical utility of their models using INR control, time in the therapeutic range, or adverse event rates.[28,44,45,48] Among these, ML-guided dosing was associated with improvements in INR target achievement and time to therapeutic range, but not reduction in adverse events. While this limited evidence suggests that ML-guided dosing may offer clinically meaningful benefits, these findings require confirmation in more robust, well-designed studies.

Implications for Practice and Policy

Our findings suggest that ML-based warfarin dosing models generally achieve higher predictive accuracy than conventional algorithms, potentially improving clinical outcomes such as achieving INR targets more quickly. However, current literature indicates that most models lack external validation and clinical implementation studies. In practice, this means that while ML tools show promise, they are not yet ready for adoption in routine clinical use without further validation in diverse populations and prospective settings. That causes the big gap between research and the ability to apply in real-world settings.

From a policy perspective, stakeholders and regulatory bodies should act based on findings from high-quality, well-designed studies with transparent reporting, using standardized outcome metrics to evaluate and compare the performance of ML models. Integration of ML dosing tools into electronic medical records or clinical decision support systems may also be considered, but only after sufficient real-world evidence has been generated.

From a research perspective, future studies are encouraged to address the methodological issues identified in this systematic review, particularly by incorporating external validation of the ML models. The study design and reporting outcomes need to ensure transparency and adhere to the reporting standards as proposed by Stevens (2020).[52] In addition, prospective trials evaluating predictive accuracy and patient-centered outcomes are strongly encouraged.

Comparison with Previous Studies

First, this study is one of the most comprehensive reviews to date, as of December 2024, of the application of various ML models to warfarin dosing in clinical settings. Thirty-five studies published between 2004 and 2024 were included. This expands upon recent works, such as Fülöp et al. (2025, which included 14 studies)[53] and Alrabadi et al. (2026, included 17 studies).[54]

Second, this review provides a distinct evaluation of clinical utility. Unlike previous reviews, we categorized and synthesized performance metrics into three distinct domains: predictive accuracy, model fit, and clinical utility. While Li et al. (2024)[6] only listed metrics without numerical synthesis, and Alrabadi et al. (2026)[54] focused solely on MAE, our review highlights the clinical utility gap. This is important because high predictive accuracy does not automatically translate to clinical safety or effectiveness.

Lastly, we summarized the evidence on the direct comparisons of the effectiveness of ML-based models versus conventional methods (empirical dosing or non-ML regression). This highlights the relative advantages and limitations of ML in real-world scenarios, offering evidence-based insights for future implementation studies.

Strengths and Limitations of this Study

This review has several strengths. First, it synthesized existing evidence on comparative performance, including predictive accuracy, model fit, and clinically relevant metrics, of ML models versus conventional methods for predicting warfarin doses in clinical settings, highlighting the potentially superior performance of ML-based approaches. Second, we employed a systematic search strategy across three major databases and adhered to PRISMA guidelines to ensure transparency and reproducibility. Third, the use of PROBAST provided a structured, standardized risk of bias assessment.

Nevertheless, several limitations should be acknowledged. All included studies were retrospective cohort studies, and external validation was limited, reducing confidence in their clinical applicability and generalizability. Substantial heterogeneity in patient populations, algorithms used, predictors, and reported outcome metrics prevented quantitative synthesis or meta-analysis. Furthermore, publication bias cannot be ruled out, as studies with null or negative results may have been underreported. These limitations limit the generalizability of our conclusions and underscore the need for more rigorous prospective validation and real-world implementation studies.

While ML models show significant promise for optimizing warfarin therapy, the current evidence base is limited by methodological weaknesses and a lack of clinical outcome validation. Future studies should prioritize prospective validation, integration with clinical workflows, and real-world outcome evaluation to assess the true value of ML in warfarin dosing.[19,40] Standardizing performance metrics and promoting transparent reporting practices are essential for building a reliable evidence base. Ultimately, collaborative efforts among clinicians, data scientists, and regulatory bodies will be crucial for translating these powerful predictive tools into safe and effective clinical practice.

Conclusion

Although ML models demonstrated generally superior performance for warfarin dosing compared with conventional methods, their clinical readiness was significantly hampered by a high risk of bias, a lack of external validation, and a scarcity of clinically relevant outcomes in the existing literature. Rigorous study designs, external validation, and measurement of clinically relevant outcomes in future studies are required to bridge the gap between algorithm development and safe, effective clinical implementation.

Authors' Contribution

All authors conceptualized the study. R.N., H.N., K.L.T.D., T.N., and K.N.C.D. designed, planned, and oversaw the study. R.N., H.N., and T.N. did a pilot search. R.N. and H.N. comprehensively searched the literature. R.N., H.N., K.L.T.D., and T.N. applied eligibility criteria and extracted data, and H.N. and N.N. conducted the risk of bias assessment. R.N., H.N., and K.N.C.D. synthesized the data. R.N., H.N., and K.N.C.D. wrote the first draft of the manuscript. All authors contributed to the interpretation, reviewed, and commented on the manuscript draft, and approved it for submission. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Data Availability Statement

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The support received for this work was an institutional research grant from DIPA PNBP Universitas Udayana TA-2025, not salary, honorarium, or personal fees to any individual author.

Ethical Approval

This study did not require review by an ethics committee as it is a systematic review based exclusively on previously published and publicly available data, with no involvement of human participants or identifiable personal information. In accordance with journal requirements, the institutional contact for matters related to research integrity is: Research and Community Services Institute (LPPM), Udayana University (e-mail: info-lppm@unud.ac.id).

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by an institutional research grant from DIPA PNBP Universitas Udayana TA-2025 under the Research Assignment Agreement Letter (Number: B/229.259/UN14.4.A/PT.01.03/2025, dated April 28, 2025). The funding represented institutional grant support only and did not constitute personal fees, salary, or honorarium to any individual author. The funders had no role in the conduct of this study or in the collection, management, analysis, and interpretation of the data. The content is solely the responsibility of the authors and does not necessarily represent the official views of Udayana University.

Informed Consent

Not applicable.

Use of Artificial Intelligence-assisted Tools

Gemini and ChatGPT were used as proofreading tools during the editing phase of this work to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Supplemental Material

Supplemental material for this article is available online.

References

1. Patel S, Singh R, Preuss CV, et al. Warfarin. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2025.
2. Pirmohamed M. Warfarin: Almost 60 years old and still causing problems. *Br J Clin Pharmacol* 2006; 62(5): 509–511. DOI: 10.1111/j.1365-2125.2006.02806.x.
3. Amaraneni A, Chippa V, Goldin J, et al. Anticoagulation safety. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2025.
4. Klein TE, Altman RB, Eriksson N, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med* 2009; 360(8): 753–764. DOI: 10.1056/NEJMoa0809329.
5. Gage BF, Eby C, Johnson JA, et al. Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. *Clin Pharmacol Ther* 2008; 84(3): 326–331. DOI: 10.1038/clpt.2008.10.

6. Li QY, Tang BH, Wu YE, et al. Machine learning: A new approach for dose individualization. *Clin Pharmacol Ther* 2024; 115(4): 727–744. DOI: 10.1002/cpt.3049.
7. Zhang F, Liu Y, Ma W, et al. Nonlinear machine learning in warfarin dose prediction: Insights from contemporary modelling studies. *J Pers Med* 2022; 12(5): 717. DOI: 10.3390/jpm12050717.
8. Deo RC. Machine learning in medicine. *Circulation* 2015; 132(20): 1920–1930. DOI: 10.1161/circulationaha.115.001593.
9. Choi RY, Coyner AS, Kalpathy-Cramer J, et al. Introduction to machine learning, neural networks, and deep learning. *Transl Vis Sci Technol* 2020; 9(2): 14. DOI: 10.1167/tvst.9.2.14.
10. Hu J and Szymczak S. A review on longitudinal data analysis with random forest. *Brief Bioinform* 2023; 24(2): bbad002. DOI: 10.1093/bib/bbad002.
11. Jiang T, Gradus JL and Rosellini AJ. Supervised machine learning: A brief primer. *Behav Ther* 2020; 51(5): 675–687. DOI: 10.1016/j.beth.2020.05.002.
12. Handelman GS, Kok HK, Chandra RV, et al. eDoctor: Machine learning and the future of medicine. *J Intern Med* 2018; 284(6): 603–619. DOI: 10.1111/joim.12822.
13. Iancu A, Leb I, Prokosch HU, et al. Machine learning in medication prescription: A systematic review. *Int J Med Inform* 2023; 180: 105241. DOI: 10.1016/j.ijmedinf.2023.105241.
14. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 2021; 372: n71. DOI: 10.1136/bmj.n71.
15. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med* 2019; 170(1): W1–W33. DOI: 10.7326/M18-1377.
16. Efthimiou O, Seo M, Chalkou K, et al. Developing clinical prediction models: A step-by-step guide. *BMJ* 2024; 386: e078276. DOI: 10.1136/bmj-2023-078276.
17. Liu R, Li X, Zhang W, et al. Comparison of nine statistical model based warfarin pharmacogenetic dosing algorithms using the racially diverse International Warfarin Pharmacogenetic Consortium cohort database. *PLoS One* 2015; 10(8): e0135784. DOI: 10.1371/journal.pone.0135784.
18. Tao Y, Chen YJ, Xue L, et al. An ensemble model with clustering assumption for warfarin dose prediction in Chinese patients. *IEEE J Biomed Health Inform* 2019; 23(6): 2642–2654. DOI: 10.1109/jbhi.2019.2891164.
19. Choi H, Kang HJ, Ahn I, et al. Machine learning models to predict the warfarin discharge dosage using clinical information of inpatients from South Korea. *Sci Rep* 2023; 13(1): 22461. DOI: 10.1038/s41598-023-49831-6.
20. Solomon I, Maharshak N, Chechik G, et al. Applying an artificial neural network to warfarin maintenance dose prediction. *Isr Med Assoc J* 2004; 6(12): 732–735.
21. Cosgun E, Limdi NA and Duarte CW. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. *Bioinformatics* 2011; 27(10): 1384–1389. DOI: 10.1093/bioinformatics/btr159.
22. Hu YH, Wu F, Lo CL, et al. Predicting warfarin dosage from clinical data: A supervised learning approach. *Artif Intell Med* 2012; 56(1): 27–34. DOI: 10.1016/j.artmed.2012.04.001.
23. Grossi E, Podda GM, Pugliano M, et al. Prediction of optimal warfarin maintenance dose using advanced artificial neural networks. *Pharmacogenomics* 2014; 15(1): 29–37. DOI: 10.2217/pgs.13.212.
24. Saleh MI and Alzubiedi S. Dosage individualization of warfarin using artificial neural networks. *Mol Diagn Ther* 2014; 18(3): 371–379. DOI: 10.1007/s40291-014-0090-7.
25. Li X, Liu R, Luo ZY, et al. Comparison of the predictive abilities of pharmacogenetics-based warfarin dosing algorithms using seven mathematical models in Chinese patients. *Pharmacogenomics* 2015; 16(6): 583–590. DOI: 10.2217/pgs.15.26.
26. Sharabiani A, Bress A, Douzali E, et al. Revisiting warfarin dosing using machine learning techniques. *Comput Math Methods Med* 2015; 2015: 560108. DOI: 10.1155/2015/560108.
27. Alzubiedi S and Saleh MI. Pharmacogenetic-guided warfarin dosing algorithm in African-Americans. *J Cardiovasc Pharmacol* 2016; 67(1): 86–92. DOI: 10.1097/fjc.0000000000000317.
28. Pavani A, Naushad SM, Kumar RM, et al. Artificial neural network-based pharmacogenomic algorithm for warfarin dose optimization. *Pharmacogenomics* 2016; 17(2): 121–131. DOI: 10.2217/pgs.15.161.
29. Li Q, Tao H, Wang J, et al. Warfarin maintenance dose prediction for patients undergoing heart valve replacement: A hybrid model with genetic algorithm and back-propagation neural network. *Sci Rep* 2018; 8(1): 9712. DOI: 10.1038/s41598-018-27772-9.
30. Ma Z, Wang P, Gao Z, et al. Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose. *PLoS One* 2018; 13(10): e0205872. DOI: 10.1371/journal.pone.0205872.
31. Roche-Lima A, Roman-Santiago A, Feliu-Maldonado R, et al. Machine learning algorithm for predicting warfarin dose in Caribbean Hispanics using pharmacogenetic data. *Front Pharmacol* 2020;10:1550. DOI: 10.3389/fphar.2019.01550.
32. Tao Y, Chen YJ, Fu X, et al. Evolutionary ensemble learning algorithm to modeling of warfarin dose prediction for Chinese. *IEEE J Biomed Health Inform* 2019; 23(1): 395–406. DOI: 10.1109/jbhi.2018.2812165.
33. Li Q, Wang J, Tao H, et al. The prediction model of warfarin individual maintenance dose for patients undergoing heart valve replacement, based on the back propagation neural network. *Clin Drug Investig* 2020; 40(1): 41–53. DOI: 10.1007/s40261-019-00850-0.
34. Huang Q, Cao L, Luo N, et al. Predicting range of initial warfarin dose based on pharmacometabolomic and genetic

inputs. *Clin Pharmacol Ther* 2021; 110(6): 1585–1594. DOI: 10.1002/cpt.2407.

35. Liu Y, Chen J, You Y, et al. An ensemble learning based framework to estimate warfarin maintenance dose with cross-over variables exploration on incomplete data set. *Comput Biol Med* 2021; 131: 104242. DOI: 10.1016/j.compbiomed.2021.104242.
36. Nguyen VL, Nguyen HD, Cho YS, et al. Comparison of multivariate linear regression and a machine learning algorithm developed for prediction of precision warfarin dosing in a Korean population. *J Thromb Haemost* 2021; 19(7): 1676–1686. DOI: 10.1111/jth.15318.
37. Steiner HE, Giles JB, Patterson HK, et al. Machine learning for prediction of stable warfarin dose in US Latinos and Latin Americans. *Front Pharmacol* 2021; 12: 749786. DOI: 10.3389/fphar.2021.749786.
38. Tao Y, Jiang B, Xue L, et al. Evolutionary synthetic oversampling technique and cocktail ensemble model for warfarin dose prediction with imbalanced data. *Neural Comput Appl* 2021; 33(17): 11203–11221. DOI: 10.1007/s00521-020-05568-1.
39. Truda G and Marais P. Evaluating warfarin dosing models on multiple datasets with a novel software framework and evolutionary optimisation. *J Biomed Inform* 2021; 113: 103634. DOI: 10.1016/j.jbi.2020.103634.
40. Amruthlal M, Devika S, Krishnan V, et al. Development and validation of a mobile application based on a machine learning model to aid in predicting dosage of vitamin K antagonists among Indian patients post mechanical heart valve replacement. *Indian Heart J* 2022; 74(6): 469–473. DOI: 10.1016/j.ihj.2022.10.002.
41. Asiimwe IG, Blockman M, Cohen K, et al. Stable warfarin dose prediction in sub-Saharan African patients: A machine-learning approach and external validation of a clinical dose-initiation algorithm. *CPT Pharmacometrics Syst Pharmacol* 2022; 11(1): 20–29. DOI: 10.1002/psp4.12740.
42. Gu ZC, Huang SR, Dong L, et al. An adapted neural-fuzzy inference system model using preprocessed balance data to improve the predictive accuracy of warfarin maintenance dosing in patients after heart valve replacement. *Cardiovasc Drugs Ther* 2022; 36(5): 879–889. DOI: 10.1007/s10557-021-07191-1.
43. Maghsoudi R, Mirzarezaee M, Sadeghi M, et al. Determining the adjusted initial treatment dose of warfarin anticoagulant medicine using kernel-based support vector regression. *Comput Methods Programs Biomed* 2022; 214: 106589. DOI: 10.1016/j.cmpb.2021.106589.
44. Zeng J, Shao J, Lin S, et al. Optimizing the dynamic treatment regime of in-hospital warfarin anticoagulation in patients after surgical valve replacement using reinforcement learning. *J Am Med Inform Assoc* 2022; 29(10): 1722–1732. DOI: 10.1093/jamia/ocac088.
45. Dryden L, Song J, Valenzano TJ, et al. Evaluation of machine learning approaches for predicting warfarin discharge dose in cardiac surgery patients: Retrospective algorithm development and validation study. *JMIR Cardio* 2023; 7: e47262. DOI: 10.2196/47262.
46. Jahmunah V, Chen S, Oh SL, et al. Automated warfarin dose prediction for Asian, American, and Caucasian populations using a deep neural network. *Comput Biol Med* 2023; 153: 106548. DOI: 10.1016/j.compbiomed.2023.106548.
47. Sridharan K, Ramanathan M and Al Banna R. Evaluation of supervised machine learning algorithms in predicting the poor anticoagulation control and stable weekly doses of warfarin. *Int J Clin Pharm* 2023; 45(1): 79–87. DOI: 10.1007/s11096-022-01471-y.
48. Petch J, Nelson W, Wu M, et al. Optimizing warfarin dosing for patients with atrial fibrillation using machine learning. *Sci Rep* 2024; 14(1): 4516. DOI: 10.1038/s41598-024-55110-9.
49. Wang M, Qian Y, Yang Y, et al. Improved stacking ensemble learning based on feature selection to accurately predict warfarin dose. *Front Cardiovasc Med* 2024; 10: 1320938. DOI: 10.3389/fcvm.2023.1320938.
50. Ma W, Li H, Dong L, et al. Warfarin maintenance dose prediction for Chinese after heart valve replacement by a feedforward neural network with equal stratified sampling. *Sci Rep* 2021; 11(1): 13778. DOI: 10.1038/s41598-021-93317-2.
51. Tao H, Li Q, Zhou Q, et al. A prediction study of warfarin individual stable dose after mechanical heart valve replacement: Adaptive neural-fuzzy inference system prediction. *BMC Surg* 2018; 18(1): 10. DOI: 10.1186/s12893-018-0343-1.
52. Stevens LM, Mortazavi BJ, Deo RC, et al. Recommendations for reporting machine learning analyses in clinical research. *Circ Cardiovasc Qual Outcomes* 2020; 13(10): e006556. DOI: 10.1161/circoutcomes.120.006556.
53. Fülöp P, Tóth Š, Porubän T, et al. Machine learning for warfarin therapy: A systematic review. *Pharmaceuticals (Basel)* 2025; 18(10): 1544. DOI: 10.3390/ph18101544.
54. Alrabadi B, Marouf M, Bashaireh T, et al. Systematic review and meta-analysis of AI accuracy in warfarin dose prediction across ethnic groups. *Eur J Clin Pharmacol* 2026; 82(2): 55. DOI: 10.1007/s00228-025-03989-2.