Introduction

Warfarin is an anticoagulant drug that has been used to treat and prevent blood clots such as venous thrombosis and pulmonary embolism. This drug works by inhibiting the activity of vitamin K epoxide reductase enzyme (Nguyen *et al*., 2021), which leads to the depletion of vitamin K-dependent clotting factors (Jahmunah *et al*., 2023), thus reducing the formation of blood clot (Rodríguez-Fernández *et al*., 2024). Even though warfarin has been known for its effectiveness, establishing the correct dosage is challenging due to its narrow therapeutic index and the wide inter-individual variability in its pharmacokinetics and pharmacodynamics (Jahmunah *et al*., 2023). Furthermore, non-optimal dosing significantly increases the risk of thromboembolism and bleeding (Xue *et al*., 2024).

CYP2C9 and VKORC1 are key genes influencing warfarin response. VKORC1 encodes the target enzyme of warfarin while CYP2C9 is responsible for its metabolism (Rangaraj & Ankani, 2024). In this project, single nucleotide polymorphism involved for both genes are VKORC1 (rs9923231) and CPY2C9 (rsrs1799853 and rs1057910).

This study aimed to develop a linear regression model that could predict the optimal warfarin dose by incorporating both clinical and genetic factors. Through this model, its predictive performance was evaluated and the relative importance of each factor in determining dosage was identified. Ultimately, the findings of this study provided deeper insight into how combined clinical and genetic data could be used to refine warfarin therapy, paving the way for more personalized medicine.

The dataset used in this project was derived from the International Warfarin Pharmacogenetics Consortium (IWPC). While the IWPC originally developed a large-scale, clinically validated dosing algorithm, the present study differs in scope and purpose. Here, the dataset was repurposed as a proof-of-concept machine learning project, focusing on data preprocessing, regression modelling, and performance evaluation. This distinction ensures that the project demonstrates technical competency in applied machine learning rather than attempting to reproduce a clinically validated algorithm.

Objectives

The primary objectives of this study were:

1. To develop a predictive model for warfarin dosing by incorporating both clinical factors (age, gender, weight, height) and genetic polymorphisms (*CYP2C9* and *VKORC1*).
2. To evaluate the performance of a machine learning approach (linear regression model) in estimating therapeutic warfarin dose.
3. To identify the relative importance of clinical versus genetic factors in predicting dose variability.
4. To compare the predicted dose with the actual prescribed dose to assess model accuracy and potential clinical applicability.

Methods

Data Collection

Patient-level data containing both genetic and clinical information were utilised for this study. The dataset included the following variables:

- Genetic factors: *CYP2C9* diplotype and *VKORC1* (rs9923231) genotype.
- Clinical factors: Age (years), gender, body weight (kg), and height (cm).
- Outcome variable: Stable therapeutic warfarin dose (mg/week).

Data Preprocessing

Data cleaning and transformation were performed before model development. Records with missing values were removed to ensure data integrity and accuracy. Age was provided in categorical ranges (e.g., “60–69 years”), which were converted into continuous values by taking the midpoint of each range. For open-ended categories (e.g., “90+”), the age was assigned as 90 years.

Genetic and categorical variables were encoded numerically:

- *CYP2C9* diplotypes were mapped as follows: *1/*1 = 0, *1/*2= 1, *1/*3 = 1, *2/*2=2, *2/*3= 2, *3/*3 = 2.
- *VKORC1* genotypes were mapped as G/G = 0, A/G = 1, A/A = 2.
- Gender was encoded as male = 1, female = 0.

All clinical continuous variables (age, weight, height) were maintained in their original numerical form.

Model Development

A linear regression model was trained to predict the therapeutic warfarin dose using both genetic and clinical factors. The model was implemented in Python (version 3.12.11) within Google Colab with the scikit-learn library. The independent variables (features) included age, gender, weight, height, *CYP2C9*, and *VKORC1*. The dependent variable (target) was the stable therapeutic dose.

Model Evaluation

The dataset was divided into training and testing subsets using an 80:20 ratio. Model performance was evaluated on the test set using the following approaches:

1. Scatter plot of predicted vs actual doses, to visually assess prediction accuracy.
2. Residual (error) distribution plot, to examine bias and variance in dose prediction.
3. Feature importance analysis was performed based on regression coefficients to identify the relative contribution of clinical and genetic predictors.
4. Performance metrics, including the coefficient of determination (R^2) and Mean Absolute Error (MAE).

All statistical analyses and visualisations were performed in Python using the pandas, numpy, matplotlib, and seaborn libraries.

Results and Discussions

Model Performance

The regression model was trained using both genetic (CYP2C9, VKORC1) and clinical variables (age, gender, weight, height). The evaluation on the test dataset showed:

- Mean Absolute Error (MAE): 8.68 mg
- Root Mean Squared Error (RMSE): 12.26 mg
- R^2 Score**:** 0.46

These metrics indicate that the model was able to predict warfarin dose with moderate to good accuracy, with an average prediction error of approximately 8.68 mg. In addition, the model achieved a coefficient of determination (R^2) of 0.46, showing that approximately 46% of the variability in warfarin dosing could be explained by the included clinical and genetic factors.

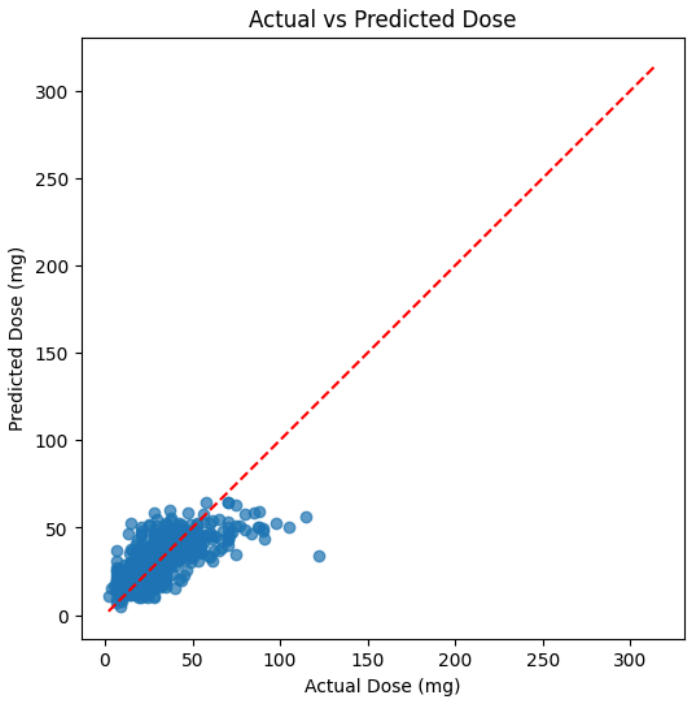

Figure 1: Graph actual vs. predicted dose

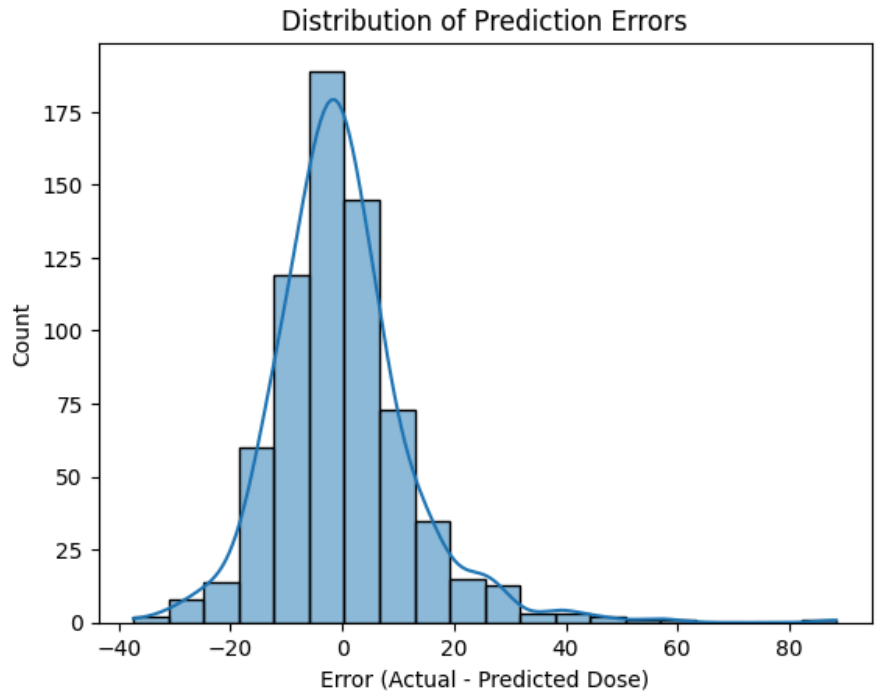

Figure 2: Graph of prediction error

The scatter plot of actual versus predicted warfarin doses in Figure 1 shows that the model performs well within low to moderate dose range (0-60 mg), with most of the data points clustered near the diagonal red line within this range. However, as the actual dose increased especially in unusually higher doses (>100 mg), the predicted value decreased, meaning that the model tends to underpredicts in the extreme range. This underprediction highlights limitations in the model's predictive accuracy to patients with unusually high dose requirements.

The histogram of prediction error plot in Figure 2 provides additional insight into the model's performance. The distribution is centred around zero, indicating that the model is generally accurate in estimating therapeutic dose and does not exhibit a strong overall bias toward over- or underestimation. However, the error distribution is slightly skewed to the positive side, which is consistent with the scatter plot observation that higher actual doses tend to be underestimated by the model. Even though there are slight deviations, the overall shape of distribution signifies that the model performance is reasonably consistent across the dataset.

Feature Importance

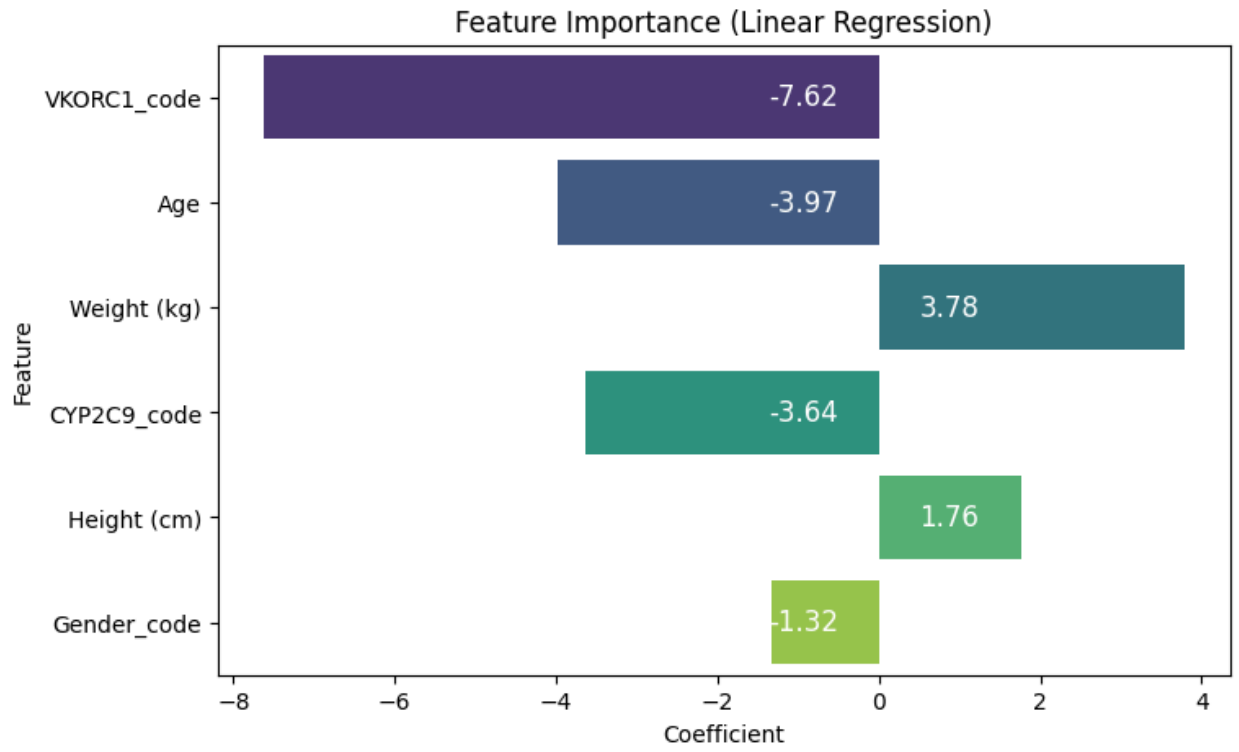

Figure 3: Graph of Feature Importance

The linear regression model was developed to predict the warfarin dosage by using both genetic and clinical factors. The regression coefficients (or feature importances) were analysed to understand which variables contributed most to dose prediction. The coefficient value indicates how much the mean of a dependent variable changes when the independent variable changes by one unit while assuming all the other factors are constant (Frost, n.d.).

Single nucleotide polymorphism (SNP) in VKORC1 which is rs9923231, is the most VKORC1 variant that related to variability in warfarin dose requirements and drug sensitivity as it has the ability to alter the promoter activity and the transcription factor binding site that lead the the reduction about 44% in the luciferase activity of the T allele compared to the C allele (Al Ammari *et al*., 2020).

VKORC1 is the most significant feature. Its large negative coefficient (-7.62) indicates it has the strongest influence of dose prediction. This inverse relationship shows that changes in VKORC1 genotype are associated with lower warfarin dose requirement. This finding is consistent with the established studies where change in VKORC1 genotype lead to an increase in the sensitivity to warfarin. Thus, low doses are required (Dean, 2012).

Genetic variants in CYP2C9 such as CYP2C9*2 (rs1799853) and CYP2C9*3 (rs1057910) are common and significantly influence warfarin dosing (Rangaraj & Ankani, 2024). Genotype CYP2C9 also has a negative coefficient even though its influence to dose prediction is slightly lower than age and weight. Changes in genotype of CYP2C9 are associated with decrease in metabolism, thus lower doses are required (Lindley *et al*., 2022). A report by Rangarah and Ankani (2024) further supports this, showing that patients carrying the wild-type genotype (CYP2C9*1) required higher warfarin doses compared to those with the variant genotype (CYP2C9*2 and CYP2C9*3).

Clinical factors have varying levels of importance. Based on the model, age appeared to be the most important clinical factors followed by weight, height and gender. Gender and height have coefficients close to zero, indicating they are the least important and have minimal impact on the model's predictions.

Age contributed more in dose prediction compared to other clinical features, its negative coefficient indicates that increasing age is associated with a lower predicted warfarin dose. This observation aligns with the findings of Fahmi *et al*. (2022), who reported that older patients generally require lower doses of warfarin compared to younger individuals.

Gender factors (Gender_code) showed negative coefficients. In this model, female was encoded as 0, while male was encoded as 1. This means that as the gender value increases, the predicted dose decreases. Therefore, the model suggests that males require lower doses compared to females. This finding is consistent with Rad *et al*. (2019), who reported that females require higher daily doses to achieve therapeutic INR compared to males. However, it contradicts the findings of Khoury & Sheikh-Taha (2014), who observed that females required a lower total weekly dose of warfarin compared to males. In addition, both studies reported no statistically significant relationship between gender and warfarin dose.

In this model, both weight and height shows positive coefficient on the model prediction.The result indicates that higher weight is associated with higher requirement for warfarin dose. This result aligned with previous findings reported by Alshammari *et al*. (2020). Moreover, increase in weight would lead to an increase in the volume of distribution and clearance of warfarin, thus the levels of coagulation factors would increase. Hence more doses of warfarin are required (Alshammari *et al*., 2020).

A study about the association of warfarin dose with the genetic and clinical factors on children with kawasaki diseases reported that height is the main factor for the prediction of warfarin dose compared to age or weight. This study also mentions that increase in height would lead to increase in warfarin dosing, maybe due to the close relationship between height and liver size (Yang *et al*., 2019).

Conclusion

The linear regression model developed in this study successfully incorporated both genetic and clinical factors to predict warfarin dosage, achieving a moderate predictive performance with an R^2 score of 0.46. The analysis of the model's performance revealed that even though it can reasonably predict doses within the typical therapeutic range, its accuracy decreases in the high-dose range, where it tends to underestimate the required dose.

To mitigate this problem, future work should include more high-dose patients in the training dataset and test non-linear models such as Random Forest or XGBoost to better capture complex dose-response patterns. Moreover, to enhance model robustness and predictive reliability, additional clinical features such as dietary factors, body mass index, comorbidities and concomitant medications should be added.

The feature importance analysis, a primary objective of this study, clearly demonstrated that genetic factors (specifically `VKORC1` and `CYP2C9`) and clinical factors (age and weight) are the main determinants of warfarin dose variability. Moreover, VKORC1 is the most dominant which aligns with established clinical literature, confirming the critical role of pharmacogenomics in personalized medicine.

Furthermore, there was an interesting finding related to the relationship between gender and warfarin dose. The model's negative coefficient for gender suggests that males require a lower dose, which aligns with some published studies but contradicts others. This highlights a key limitation of relying solely on machine learning models. Their findings, especially for less dominant features, may be influenced by data-specific correlations rather than universal clinical truths. The study underscores the importance of interpreting model results critically and validating them against a broad body of domain knowledge.

In summary, this project validates the predictive power of a simple linear model using a combination of clinical and genetic data. It successfully identifies genetic factors as the most crucial predictors of warfarin dose, confirming their clinical significance. The study also serves as an important case study on the value of integrating domain expertise to provide a complete and accurate interpretation of machine learning outputs.

References

Al Ammari, M., AlBalwi, M., Sultana, K., Alabdulkareem, I. B., Almuzzaini, B., Almakhlafi, N. S., ... & Alghamdi, J. (2020). The effect of the VKORC1 promoter variant on warfarin responsiveness in the Saudi Warfarin Pharmacogenetic (SWAP) cohort. *Scientific Reports*, *10*(1), 11613.

Alshammari, A., Altuwayjiri, A., Alshaharani, Z., Bustami, R., & Almodaimegh, H. S. (2020). Warfarin Dosing Requirement According to Body Mass Index. *Cureus*, *12*(10), e11047. https://doi.org/10.7759/cureus.11047

Dean, L. (2012). Warfarin therapy and VKORC1 and CYP2C9 genotypes. In M. P. Adam, H. H. Ardinger, R. A. Pagon, et al. (Eds.), *GeneReviews*® (pp. 1–17). University of Washington, Seattle. https://www.ncbi.nlm.nih.gov/books/NBK84174/

Fahmi, A. M., Elewa, H., & El Jilany, I. (2022). Warfarin dosing strategies evolution and its progress in the era of precision medicine, a narrative review. *International journal of clinical pharmacy, 44*(3), 599–607. https://doi.org/10.1007/s11096-022-01386-8

Frost, J. (n.d.). How to interpret P-values and coefficients in regression analysis. *Statistics by Jim*. https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

Jahmunah, V., Chen, S., Oh, S. L., Acharya, U. R., & Chowbay, B. (2023). Automated warfarin dose prediction for Asian, American, and Caucasian populations using a deep neural network. *Computers in Biology and Medicine*, *153*, 106548.

Khoury, G., & Sheikh-Taha, M. (2014). Effect of age and sex on warfarin dosing. *Clinical pharmacology : advances and applications*, *6*, 103–106. https://doi.org/10.2147/CPAA.S66776

Lindley, K. J., Limdi, N. A., Cavallari, L. H., Perera, M. A., Lenzini, P., Johnson, J. A., ... & Gage, B. F. (2022). Warfarin dosing in patients with CYP2C9* 5 variant alleles. *Clinical Pharmacology & Therapeutics*, *111*(4), 950-955.

Nguyen, H. D., Cho, Y. S., Kim, H. S., Han, I. Y., Kim, D. K., Ahn, S., & Shin, J. G. (2021). Comparison of multivariate linear regression and a machine learning algorithm developed for prediction of precision warfarin dosing in a Korean population. Journal of Thrombosis and Haemostasis, 19(7), 1676-1686.

Rad, F., Hamidpour, M., Dorgalaleh, A., & Poopak, B. (2019). The Effect of Demographic Factors and VKORC1 1639 G>A Genotypes on Estimated Warfarin Maintenance Dose in Iranian Patients Under Warfarin Therapy. *Indian journal of hematology & blood transfusion : an official journal of Indian Society of Hematology and Blood Transfusion*, *35*(1), 167–171. https://doi.org/10.1007/s12288-018-0987-0

Rangaraj, S., & Ankani, B. T. S. (2024). Warfarin Dose Maintenance Associated with CYP2C9* 2 (rs1799853) and CYP2C9* 3 (rs1057910) Gene Polymorphism in North Coastal Andhra Pradesh. *The Open Biochemistry Journal*, *18*(1).

Rodríguez-Fernández, K., Reynaldo-Fernández, G., Reyes-González, S., de Las Barreras, C., Rodríguez-Vera, L., Vlaar, C., ... & Mangas-Sanjuan, V. (2024). New insights into the role of VKORC1 polymorphisms for optimal warfarin dose selection in Caribbean Hispanic patients through an external validation of a population PK/PD model. *Biomedicine & Pharmacotherapy*, *170*, 115977.

Xue, L., Singla, R. K., He, S., Arrasate, S., González-Díaz, H., Miao, L., & Shen, B. (2024). Warfarin–A natural anticoagulant: A review of research trends for precision medication. *Phytomedicine*, *128*, 155479.

Yang, D., Kuang, H., Zhou, Y., Cai, C., & Lu, T. (2019). Height, VKORC1 1173, and CYP2C9 Genotypes Determine Warfarin Dose for Pediatric Patients with Kawasaki Disease in Southwest China. *Pediatric cardiology*, *40*(1), 29–37. https://doi.org/10.1007/s00246-018-1957-x