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 factor (Jahmunah *et al*., 2023). 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).

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/*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 (R²) 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): 8.70 mg
- Root Mean Squared Error (RMSE): 14.46 mg
- R² Score**:** 0.40

These metrics indicate that the model was able to predict warfarin dose with moderate to good accuracy, with an average prediction error of approximately 8.70 mg. In addition, the model achieved a coefficient of determination (R²) of 0.40, showing that approximately 40% of the variability in warfarin dosing could be explained by the included clinical and genetic factors.

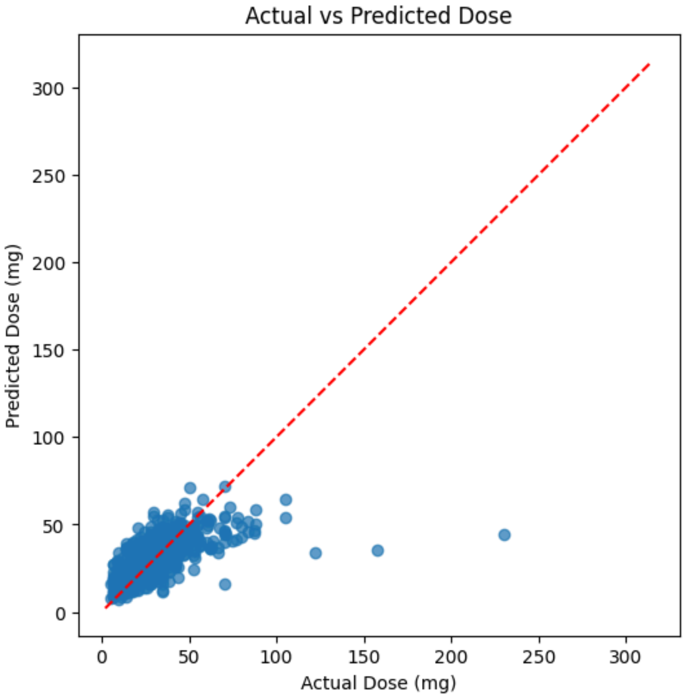

Figure 1: Graph actual vs. predicted dose

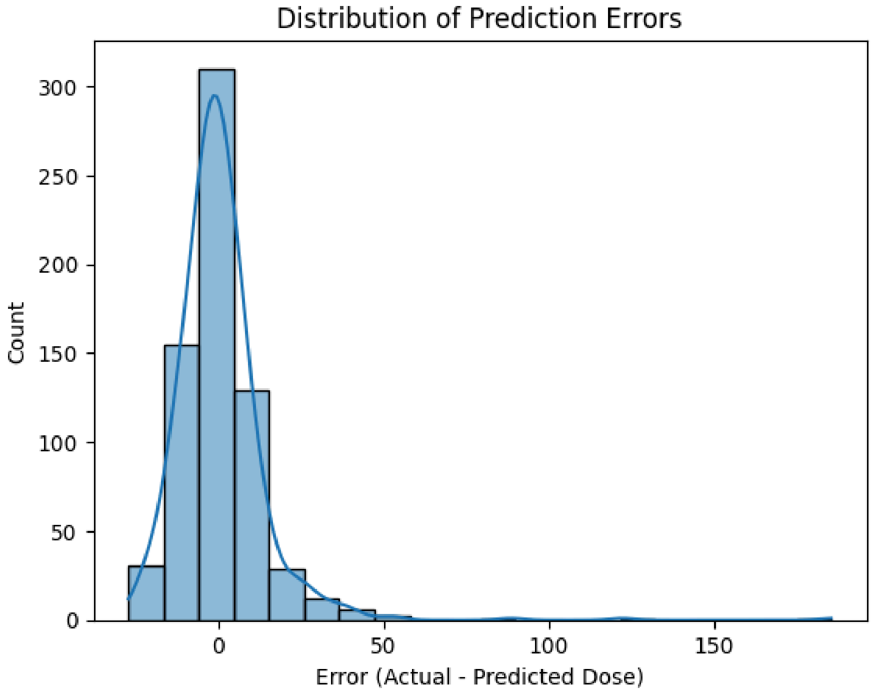

Figure 2: Graph of prediction error

The scatter plot of actual versus predicted warfarin doses in Figure 1 shows that most data points are clustered near the diagonal red line, suggesting that the model can predict the required dose. However, the model performs well only within the standard dosing range (< 75 mg/week), as most of the data points are clustered within this range. At higher dose ranges (> 100 mg/week), several scattered outliers are observed, suggesting that the model tends to underpredict in this extreme range.

The prediction error plot in Figure 2 provides additional insight into the model's performance. The distribution is centred around zero, indicating that the model 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. In addition, most errors fall within the range of –20 to +20 mg, suggesting that the model is reasonably accurate for most patients, even though prediction errors increase in those requiring very high warfarin doses.

Feature Importance

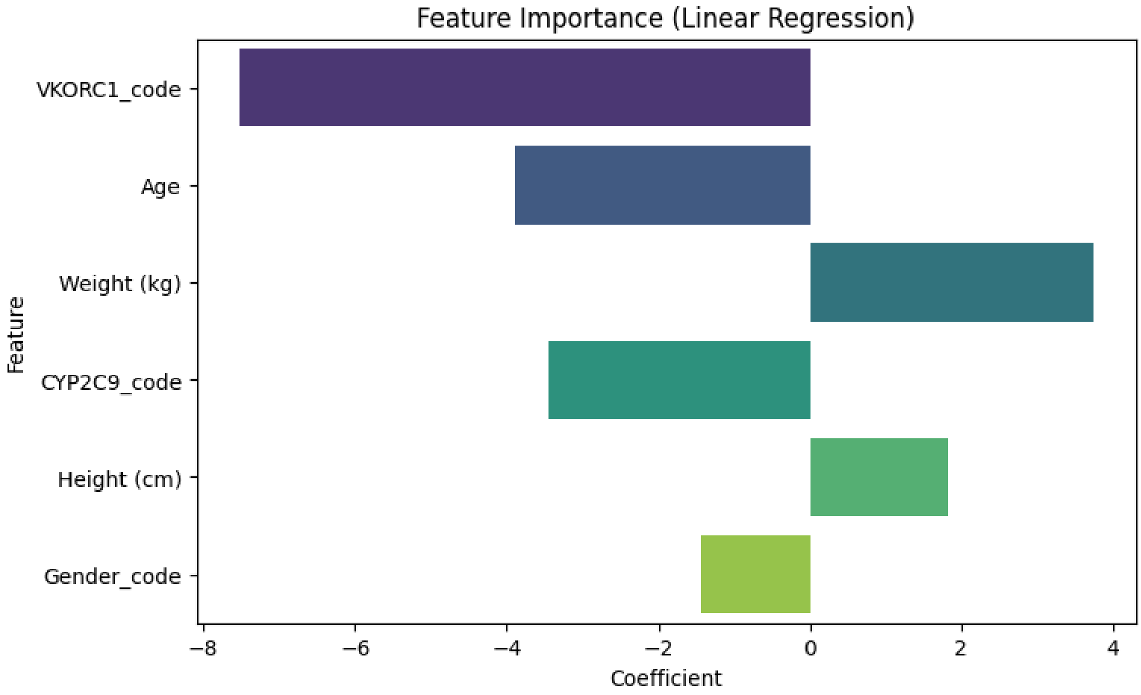

Figure 3: Graph of Feature Importance

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.).

Genetic factors(VKORC1_code and CYP2C9_code) are the most important predictors. The large coefficients for these two features show that genetic information has the strongest impact on the model's predictions. This finding aligns with reports published by Dean (2012). The report states that VKORC1 and CYP2C9 genotypes are the most important determinants of individual warfarin dose requirements.

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

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.

Although age contributed less, 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.

In this model, both weight and height shows positive coefficient even though they have minimal impact 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 0.40. 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.

The feature importance analysis, a primary objective of this study, clearly demonstrated that genetic factors (specifically `VKORC1` and `CYP2C9`) are the most dominant determinants of warfarin dose variability. This finding aligns with established clinical literature, confirming the critical role of pharmacogenomics in personalized medicine. In comparison, clinical factors such as age, weight, and height were found to have a minimal impact on the model's predictions, with coefficients close to zero.

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

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

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

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