**Title :** Development of a Clinical Decision Support Tool for Personalized Warfarin Dose Determination Using Individualized Predictive Modeling Integrating Statistical and Machine Learning Approaches

**Author(s) :**

1. Miss Keetapat Tananchai — **Student ID :** 650510694
2. Miss Thita Marayat — **Student ID :** 650510700
3. Mr. Tanakorn Yimyorng — **Student ID :** 650510706
4. Mr. Poramad Boontha — **Student ID :** 650510709
5. Miss Kanjanapron Matapan — **Student ID :** 650510723

**Major :** Data Science Program

**Advisor(s) :**

1. Assistant Professor Dr. Phimphaka Taninpong

**Type of presentation*** *(choose 1)* **:**

- ☐ **Oral Presentation** (เฉพาะ ตัวแทนนศ.ที่สาขาเลือกให้นำเสนอแบบบรรยาย)
- ☑ **Poster** (กรณี นำเสนอผลงานปัญหาพิเศษ/การค้นคว้าอิสระ)
- ☐ **Cooperative Education** (กรณี นำเสนอผลงานสหกิจศึกษา)

## ABSTRACT

This study aimed to develop, compare, and evaluate the predictive performance of machine learning models for individualized warfarin dose estimation using two data sources with distinct contextual characteristics: (1) the international dataset from the International Warfarin Pharmacogenetics Consortium (IWPC), and (2) electronic medical record data from a hospital. The study utilized a retrospective structured clinical dataset comprising demographic and clinical variables, including age, sex, body weight, height, target INR, comorbidities, and concomitant medications. The IWPC dataset additionally included pharmacogenetic variables such as *CYP2 C9 * and *VKORC1 * genotype, which are directly associated with warfarin metabolism and pharmacodynamic sensitivity. The outcome variable was the stable weekly warfarin dose (mg/week), treated as a continuous variable. Data preprocessing involved outlier detection, missing data handling, categorical encoding, feature scaling, and an 80:20 training–testing split to mitigate evaluation bias.

**Type of presentation must be matched with an option you choosing on student upload system.*

***The abstract can be more than one page and must be approved by project advisor before upload.*

This study employed 39 predictive models, encompassing both linear and nonlinear approaches, including Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, and Gradient Boosting Regression Trees (GBRT). Model ensemble strategies, such as weighted averaging and stacking-like aggregation were also implemented to examine the impact of increased model complexity and variance reduction within the bias–variance trade-off framework. Model selection was based on Argmax($R^2$) from the test set. Performance evaluation employed multiple metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination ($R^2$), Pearson's correlation coefficient (r), and Percentage Within 20% (PW20), the latter reflecting clinical dosing accuracy.

In the hospital electronic medical record test dataset, the best-performing model was GBRT_100, configured with 100 trees. It achieved MAE of 4.9111 mg/week, RMSE of 6.4922 mg/week, $R^2$ of 0.2980, and PW20 of 28.57%, with a Pearson correlation coefficient (r) of 0.694, indicating a moderately strong positive linear association between observed and predicted stable weekly warfarin doses.

In contrast, the IWPC cohort, which complete genetic variables were available, were separated into two subsets: Asian supgroup and non-Asian subgroup. For Asian subgroup, the best-performing model was GBRT_100, achieving $R^2$ of 0.2140, MAE of 6.6712 mg/week, RMSE of 9.2520 mg/week, and PW20 ≈ 40%. For Non-Asian subgroup, the GBRT_AbsoluteLoss yield the $R^2$ of 0.1968, MAE of 10.7300 mg/week, and RMSE of 14.1910 mg/week, while stacking of GBRT_AbsoluteLoss and SVR_rbf_C10_gScale model slightly improved $R^2$ of 0.2109 and RMSE of 14.0660 mg/week, with PW20 of 38.74%. In conclusion, machine learning model, particularly GBRT, demonstrate potential for predicting individualized warfarin dosing in clinical settings when sufficiently informative data are available.