공학석사 학위논문

Development of medical artificial intelligence model and construction of heterogeneous database using electronic medical record (EMR) for personalized drug treatment

전자의무기록 데이터 기반 맞춤형 약물 치료를
위한 의료 인공지능 모델 개발 및 데이터베이스
구축

울산대학교 대학원
의과학과
최희정

Development of medical artificial intelligence model and construction of heterogeneous database using electronic medical record (EMR) for personalized drug treatment

지도교수 김영학

이 논문을 공학석사학위 논문으로 제출함

2023 년 8 월

울산대학교 대학원

의과학과

최희정

최희정의 공학석사학위 논문을 인준함

심사위원 이 계 화 (인)

심사위원 김 영 학 (인)

심사위원 전 태 준 (인)

울산대학교 대학원

2023 년 8 월

Abstract

Many patients undergo drug therapy for disease treatment. However, due to the unique characteristics of each individual, even when receiving the same drug, the drug response can vary among patients. Consequently, personalized drug therapy that takes into account individual characteristics is necessary, with different drug regimens for each patient. Personalized drug therapy offers the advantage of minimizing drug side effects while maximizing treatment effectiveness. Therefore, both currently used drugs and under-development drugs should be used as personalized medications. In this paper, two studies were conducted to help personalized drug therapy by utilizing electronic medical record (EMR) data from tertiary hospitals.

In the first study, we developed and validated a machine learning model for early prediction of the discharge dosage of the anticoagulant drug warfarin. We developed four machine learning models suitable for predicting drug dosage, and through internal validation, we confirmed that the model predictions were more accurate than those of clinical experts. Additionally, we utilized the SHAP (SHapley Additive exPlanations) technique to analyze the key variables that influence the model predictions and explain the model prediction. Finally, we observed significant variability in dosage determination depend on physician's individual medical experiences, when presented with the same dataset. In contrast, the model's predictive accuracy demonstrated a clinical utility that was twice as high as those of physicians.

In the second study, we constructed a novel clinical field-based database by integrating electronic medical records (EMR) and pharmaceutical databases. FDA-approved anticancer agents and associated target gene information was extracted from the Open Targets Platform. We standardized the drug components in both the EMR and Open Targets Platform, and established a linkage between the two databases based on the drugs. As a result, the novel database was included associations between 57 anticancer agents, 60 types of cancer, and 91 genetic mutations. Besides, the database was included additional diagnostic information and genetic test results of patients prescribed with anticancer agents. This integration of data sources allowed for the utilization of both clinical and genetic characteristics of patients in real-world clinical settings, utilizing for personalized cancer treatment.

In this study, we have developed two tools that utilize electronic medical record (EMR) data to facilitate personalized drug treatment. First, the machine learning models that predicts the optimal dosage of warfarin can be used as a clinical decision support system to reduce unnecessary treatment duration and contribute to the prevention of drug side effects. Second, the heterogeneous database that integrated both of EMR data and pharmaceutical information databases can be utilized in artificial intelligence-driven drug development.

Abbreviations

EMR: Electronic Medical Record

AI: Artificial Intelligence

ML: machine learning

EHR: Electronic Health Record

MAE: Mean Absolute Error

INR: International Normalized Ratio

MIMIC-III: Medical Information Mart for Intensive Care III

ICU: Intensive Care Unit

ICD-10: International Classification of Diseases, Tenth Revision

ANN: Artificial Neural Network

XGBoost: Extreme Gradient Boosting

SHAP: Shapley Additive Explanations

ICC: intraclass correlation coefficient

RWE: Real World Evidence

ADMET: Absorption, Distribution, Metabolism, Excretion, Toxicity

SNV: Single Nucleotide Variant

MPNST: Malignant Peripheral Nerve Sheath Tumor

DTI: Drug-Target Interaction

DDI: Drug-Drug Interaction

Contents

Chapter 1.

Machine learning models to predict the warfarin discharge dosage using clinical information of East Asian inpatients

Abstract

Background: As warfarin has a narrow therapeutic window and obvious response variability among individuals, it is difficult to rapidly determine personalized warfarin dosage. Adverse drug events resulting from warfarin overdose can be critical. Our study aimed to develop a machine learning (ML) model that predicts the appropriate discharge dosage of warfarin using electronic medical records from a large hospital. Additionally, we externally validated the model to ensure its accuracy.

Objective: This study aimed to develop a machine learning model that predicts individual warfarin dosage based on clinical data within 2 days of hospitalization.

Methods: During this retrospective study, adult patients who were prescribed warfarin at a large hospital between January 1, 2018, and October 31, 2020, were recruited as a model development cohort (n=3,168). We externally validated the models using the Medical Information Mart for Intensive Care III (n=891). Variables for the warfarin dosage prediction were selected according to the clinical practice experience of cardiovascular physicians. The study outcome was the warfarin discharge dosage. Four ML models that predicted the proper warfarin discharge dosage were developed. We evaluated the model performance using the mean absolute error (MAE) and prediction accuracy. Finally, we compared the accuracy of the predictions of our models and the predictions of physicians to determine their clinical relevance.

Results: The MAEs obtained using the internal validation set were as follows: XGBoost, 0.9; artificial neural network, 1.0; random forest, 1.0; linear regression, 1.0; and physicians, 1.3. These values showed that our models had better prediction accuracy than the physicians, who have difficulty determining the warfarin discharge dosage using clinical information obtained within 2 days of hospitalization.

Conclusions: Our ML model could help physicians rapidly predict and decide the proper warfarin discharge dosage during hospitalization. Further work is required to determine model generalizability.

Keywords: warfarin; EMR; electronic medical record; EHR; electronic health record; machine learning; deep learning; dosing algorithm;

1. Introduction

Warfarin is an oral anticoagulant; it has been used for the treatment and prevention of thromboembolic disorders for more than 60 years [1]. Despite its well-studied clinical pharmacology, high efficacy, and cost-effectiveness, it is clinically challenging to 1 determine the appropriate dosage of warfarin for each individual because of its narrow therapeutic window and variable patient responses [2]. Conventionally, the international normalized ratio (INR) blood coagulation test is performed to achieve optimal efficacy and minimize side effects of warfarin, and physicians adjust the warfarin dosage individually based on their medical experience and INR values [3]. However, if the dosage is insufficient, the risk of thrombosis increases; conversely, if the dosage is excessive, then the risk of bleeding increases [4]. Warfarin is one of the ten main anticoagulants that cause adverse drug events [4]. Recent studies have attempted to predict ideal warfarin dosage using machine learning algorithms, which mostly develop models using genetic data obtained through genetic testing [5–12]. Genetic variations in CYP2C9, VKORC1, and CYP4F2 have been shown to have significant correlations with warfarin. However, genetic testing is not performed in actual clinical settings because it is time-consuming [13]. In almost of tertiary hospitals from South Korea, the genetic test takes about 2 weeks until getting the results, because there are many patients performed the genetic test and waiting the results. Consequently, genetic test is not appropriate for some patients who need a prescription of drugs immediately. Therefore, more clinically relevant studies that predict the warfarin dosage using only clinical features are required. One study aimed to predict the adjustment dose of warfarin using only clinical data [14]. However, warfarin adjustment doses are prescribed to outpatients and can interact with some foods or alcohol, thus affecting their lifestyle. In such cases, it is questionable whether using lifelog data reflecting the lifestyle and environment of patients is more appropriate than using clinical data in the electronic medical records. Accordingly, we concluded that a tool that can provide reliable predictions using only clinical data obtained from inpatients instead of complex genetic testing data is necessary. In this study, we developed machine learning models that predict the appropriate warfarin discharge dosage using only clinical data generated in hospitals within 2 days of hospitalization. Our artificial intelligence approach has the potential to improve clinical utility by rapidly presenting appropriate warfarin dosing information so that the appropriate decisions can be made for critically ill patients in clinical practice. Additionally, these machine learning models could be beneficial to both hospitals and patients by decreasing unnecessary treatment durations, thereby securing space for other patients in hospital wards and reducing the financial burden of hospitalization costs for patients.

2. Methods

Ethical approval

This study obtained approval and waived the written informed consent from the Institutional Review Boards of Asan Medical Center (No. 2021-0321). All experiments were performed in accordance with relevant guidelines and regulations

2.1 Study Design

We designed four machine learning models to predict the warfarin dosage prescribed at discharge using only clinical data measured on the first and second days of hospitalization (**Fig. 1**). We extracted clinical variables in electronic medical records (EMRs) of the Asan Medical Center (AMC) and Medical Information Mart for Intensive Care III (MIMIC-III) database [15] with same criteria for model development and validation. Then, data pre-processing was performed and the models were trained using training set. Next, the actual warfarin dosage prescribed by the physicians was setted as baseline and compared with the predictions of the models to examine that our models' predictions were more accurate and rapid than the physicians. Model performance was evaluated using both of the mean absolute error (MAE) [16], which has well-known performance metric in regression models and predictive accuracy. In addition, we externally validated the models to external validation set from the MIMIC-III. Data pre-processing, model development, training and validation were conducted in Python 3.8.10. Finally, we analyzed the predictions of five physicians by calculating intraclass correlation coefficient (ICC) [17] value and compared the predictions of the models and those of the physicians using 40 data points to explore the clinical utility of the models.

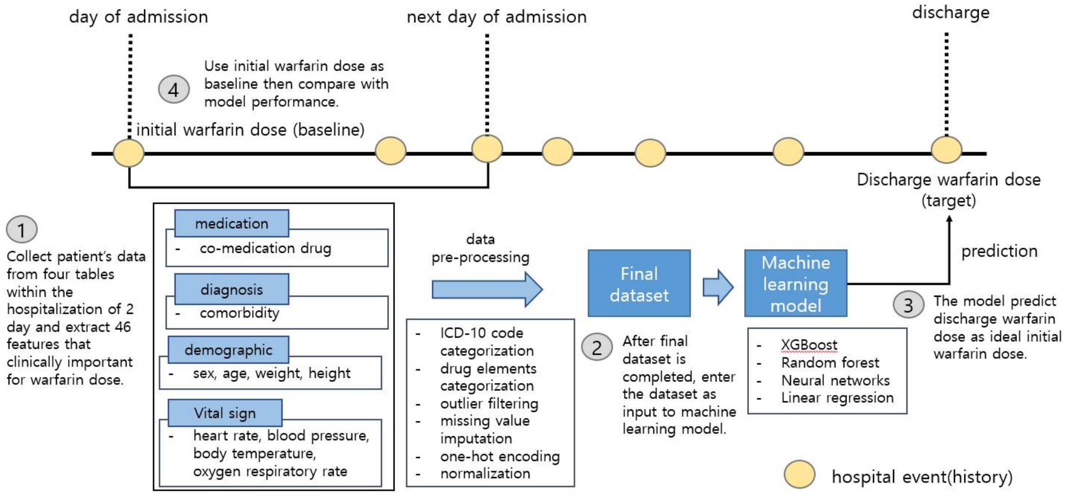
day of admission
next day of admission
discharge
4
Use initial warfarin dose as baseline then compare with model performance.
initial warfarin dose (baseline)
Discharge warfarin dose (target)
1
Collect patient's data from four tables within the hospitalization of 2 day and extract 46 features that clinically important for warfarin dose.
medication
- co-medication drug
diagnosis
- comorbidity
demographic
- sex, age, weight, height
Vital sign
- heart rate, blood pressure, body temperature, oxygen respiratory rate
data pre-processing
- ICD-10 code categorization
- drug elements categorization
- outlier filtering
- missing value imputation
- one-hot encoding
- normalization
Final dataset
2
After final dataset is completed, enter the dataset as input to machine learning model.
Machine learning model
- XGBoost
- Random forest
- Neural networks
- Linear regression
prediction
3
The model predict discharge warfarin dose as ideal initial warfarin dose.
hospital event(history)

Figure 1. Overview of the workflow. Our machine learning (ML) models complied with the workflow in the following order: (1) through discussions with clinical experts, we collected 46 features from the electronic medical record (EMR) database up to the second day of hospitalization; (2) we conducted data pre-processing, such as missing value imputation, outlier filtering, and normalization, and created the final dataset; (3) the models predicted the warfarin discharge dosage using the dataset; and (4) we compared the initial warfarin dosage with model predictions to confirm that our models could rapidly predict more accurate discharge dosage than the physicians.

2.2 Data collection

The model development cohort consisted of patients admitted to the cardiovascular or thoracic and cardiovascular surgery departments of AMC between January 1, 2018, and October 31, 2020. All the selected participants were at least 19 years old; exclusions were based on the following criteria: none of warfarin prescription at discharge; <3 warfarin prescriptions; and no weight measurements within 2 days of hospitalization (**Fig. 2**). The external validation set derived from the MIMIC-III followed the same workflow except for medical department codes; this is because the medical department codes of the intensive care unit (ICU) could not be found. Finally, the development cohort derived from AMC EMRs comprised 3,168 patients and the external validation cohort derived from the MIMIC-III comprised 891 patients.

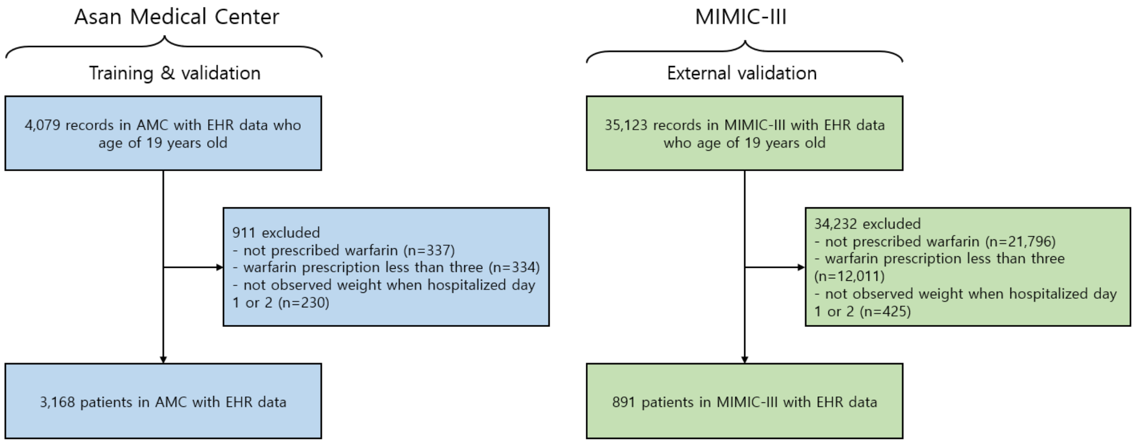
Asan Medical Center
Training & validation
4,079 records in AMC with EHR data who age of 19 years old
911 excluded
- not prescribed warfarin (n=337)
- warfarin prescription less than three (n=334)
- not observed weight when hospitalized day 1 or 2 (n=230)
3,168 patients in AMC with EHR data
MIMIC-III
External validation
35,123 records in MIMIC-III with EHR data who age of 19 years old
34,232 excluded
- not prescribed warfarin (n=21,796)
- warfarin prescription less than three (n=12,011)
- not observed weight when hospitalized day 1 or 2 (n=425)
891 patients in MIMIC-III with EHR data

Figure 1. Cohort diagram. AMC, Asan Medical Center; EHR, electronic health record; MIMIC-III, Medical

Information Mart for Intensive Care III.

2.3 Data Preparation

2.3.1 Feature Selection

The process of feature selection used for model development and validation was conducted only when had already been proofed correlations with warfarin based on clinical rationales [18–32]. The 46 clinical variables regarded as key factors of warfarin dosage adjustments and associated with thrombus or bleeding were selected in 4 tables of demographic, diagnosis, medication, vital sign can be found in **Appendix Table 1**. First of all, 4 demographic variables of age, sex, height, weight were used. Second, comorbidity variables of 15 diseases associated with thrombus or bleeding were selected in the diagnosis table and used as diagnosis variables were shown in **Table 1**. Third, 18 medicines, affecting functions of warfarin and increasing a risk of bleeding and causing warfarin dosage adjustments when combined with warfarin, were selected in the medication table and were shown as **Table 2**. Additionally, the effect of warfarin inhibiting the coagulation of the blood can be estimated through monitoring vital signs. If coagulation of the blood was caused, the symptoms including difficulty breathing and low blood oxygen saturation, rapid breathing, and a rapid heart rate, also fever are observed. Thus, the vital sign measurement values can be utilized for warfarin dosage adjustments and 6 variables were selected as follow: heart rate, oxygen respiration, systolic blood pressure, diastolic blood pressure, body temperature, respiration rate.

	Diseases
Associated with thrombus	angina, arrhythmia, atrial fibrillation, chronic ischemic heart disease, chronic lung disease, dyslipidemia, heart failure, myocardial infarction, peripheral arterial disease, pulmonary embolism, stroke, valvular heart disease
Associated with bleeding	cancer, chronic ischemic heart disease, diabetes mellitus, hypertension, intracranial bleeding, liver disease, renal disease

Table 1. A list of diseases associated with functions or side effects of warfarin

ACE inhibitor	ADP receptor inhibitor	aldosterone antagonists
allopurinol	amiodarone	ARB
aspirin	beta-blocker	calcium channel blocker, dihydropyridine
calcium channel blocker, non-dihydropyridine	diuretic, loop	diuretic, thiazide
insulin	lipid lowering	metformin
nitrate	statin	sulfonylurea

Table 2. A list of medicines affecting a dosage of warfarin when combined with warfarin

2.3.2 Categorization

We conducted categorization of variables in comorbidity and concurrent medication twice to reduce redundancy and consider more comprehensive information. It allowed us to group variables with clinical associations together and capture their collective impact on the prediction. This approach made the machine learning models more robust exploring broader patterns and relationships in the dataset. At first, the International Classification of Diseases, Tenth Revision (ICD-10) codes, which are the diagnostic codes used in the AMC EMRs, have a hierarchical structure; therefore, they can be grouped using the first disease category code. For example, code I48 includes all diseases related to atrial fibrillation and flutter. Finally, 96 diagnostic codes based on the first three characters were included for 18 diagnosis groups. All ICD-10 codes included as diagnosis variables are can be found as Supplementary **Appendix Table 2**. Then, the medication information was also categorized into each group based on the associated ingredients of the prescribed drugs. For example, rosuvastatin, simvastatin, and atorvastatin were grouped as statins, whereas cilazapril, ramipril, fosinopril, and others were classified as angiotensin-converting enzyme inhibitors. Consequently, a total of 290 medicinal components were assigned into 18 groups with similar drug effects.

2.3.3 Data Transformation

The categorical variables need to convert into numerical values because it can't use in machine learning models. Consequently, we performed one-hot encoding at three times to variables in comorbidity, concurrent medication, sex. First, the sex code underwent one-hot encoding, with 1 representing male sex and 2 representing female sex. Next, comorbidity and drug variables underwent one-hot encoding as 1 if a specific diagnostic code was assigned or a specific drug was administered within 2 days of hospitalization and as 0 if otherwise. Finally, the entire of categorical variables were transformed as numerical vectors can be used in machine learning model.

2.3.4 Imputation & Normalization

We performed data cleaning each other models, because there is difference of suitable methods with model types that planned to develop in this study. The XGBoost and Random Forest models are a decision-tree-based ensemble machine learning algorithm, but artificial neural network and linear regression models are not. Then, we performed two imputation methods to preprocess missing values in continuous variables, by considering whether the model was a decision-tree-based model. Because the tree-based models can automatically handle missing values and are not sensitive to missing values or outliers [33], any missing data was replaced with minus 1 in XGBoost and Random Forest models. Mean imputation would be appropriate to deal with the missing data for artificial neural network and linear regression models, because the continuous variables used in this study have not a wide range of values. Accordingly, we performed mean imputation that convert missing values as mean values in the specific variable in artificial neural network and linear regression models. Finally, we conducted normalization in artificial neural network and linear regression models and all variables were normalized using the minimum-maximum scaling method [34], resulting in a range from 0 to 1.

2.3.5 Data Split

The AMC dataset (n=3,168) was separated as 80% for the training set (n=2,534) and 20% for the internal validation set (n=634) using random split method. Additionally, we performed external validation to evaluate the generalizability of the model to the external validation set (n=891) from the MIMIC-III.

2.4 Model Development and Validation

The following models were developed: artificial neural network (ANN) [35], linear regression [36], extreme gradient boosting (XGBoost) [37], and random forest [38] models. These models were trained in training set including 46 clinical features and predicted a warfarin discharge dosage. We conducted Grid Search [39] with random shuffles of 5-fold cross-validation [40] to identify the optimal hyperparameters for each model. The entire of final hyperparameters of four models was shown in **Appendix Table 3**. Finally, the XGBoost and random forest models utilized the raw dataset as the input, Whereas the ANN and linear regression models used the minimum-maximum scaled dataset as the input to help a rapid optimization of each models.

2.5 Performance Metrics

We used both of MAE and predictive accuracy as performance metrics to evaluate the model prediction ability. Baseline was set to the initial dosage of warfarin prescribed by a physician upon hospitalization. The performance of the models was compared with the baseline to examine that our models' predictions were more accurate and rapid than the physicians. Subsequently, we calculated the accuracy of the model prediction using three thresholds: 0.5 mg, 1.0 mg, and 1.5 mg. This approach was consulted a logistic regression that conducts a binary classification whether the prediction was greater than a specific cut-off value [41]. The prediction was classified as accurate if the MAE of the sample was smaller than the corresponding threshold; otherwise, it was classified as inaccurate. For example, when the threshold was 0.5 mg, if the MAE of a particular sample was 0.3 mg, the prediction was classified as accurate. We calculated the proportion of samples with accurate predictions determined by each model and evaluated the accuracy of predictions based on each threshold.

2.6 Model Interpretations

We used the Shapley additive explanations (SHAP) method to obtain insights of the predictions of our models and understand how each variable contributes to predictions [42]. The SHAP method is an explainable artificial intelligence method that decomposes the output of the model into the contributions of each feature, allowing for an analysis of the influence of each feature on the model [43]. It considers dependencies between features and can calculate positive and negative impacts, unlike traditional variable importance measures. Higher SHAP values indicate that the patient needs higher warfarin dose. The SHAP values calculated using the internal validation set were applied to visualize dependence, waterfall, and beeswarm plots.

2.7 Comparison of Models' and Physicians' Predictions

We selected 20 data points with accurate model predictions, high physician prediction errors and 20 data points with high model prediction errors and accurate physician predictions. The XGBoost model was used. Subsequently, we constructed a dataset with 50% model accuracy and 50% baseline accuracy. Next, we distributed these datasets to five physicians and asked them to predict the appropriate warfarin discharge dosage. We analyzed intraclass correlation coefficient (ICC) of the physicians' predictions to test the interrater agreement using 2-way random effects model in R. Finally, we compared the predictions of the machine learning models and those of the physicians.

3. Results

3.1 Patient Characteristics

The baseline characteristics of the two datasets used for model development and validation are listed in **Table 3**.

	AMC dataset (n=3,168)	**MIMIC-III dataset (n=891)**
Demographics		
Age, mean (SD), years	62.3 (12.5)	65.3 (14.3)
Male	1,674 (52.8%)	510 (57.2%)
Female	1,494 (47.2%)	381 (42.8%)
Height, mean (SD)	162.2 (9.5)	171.0 (10.7)
Weight, mean (SD)	63.5 (13.0)	87.6 (46.1)
Vital Signs		
Heart rate, mean (SD)	70.9 (23.7)	84.5 (14.8)
O_2 saturation, mean (SD)	97.8 (2.3)	100.0 (0.7)
Systolic blood pressure, mean (SD)	115.9 (20.3)	115.0 (22..3)
Diastolic blood pressure, mean (SD)	67.4 (12.2)	58.6 (1..4)
Body temperature, mean (SD)	36.6 (0.6)	36.7 (0.7)
Respiration rate, mean (SD)	17.7 (3.0)	18.6 (3.8)
Comorbidity (n, %)		
Angina	114 (3.6%)	31 (3.5%)
Arrhythmia	172 (5.4%)	156 (17.5%)
Atrial fibrillation	1,335 (42.1%)	541 (59.6%)
Cancer	172 (5.4%)	55 (6.2%)
Chronic ischemic heart disease	368 (11.6%)	323 (36.3%)
Chronic lung disease	54 (1.7%)	210 (23.6%)
Diabetes mellitus	501 (15.8%)	295 (33.1%)
Dyslipidemia	70 (2.2%)	362 (40.6%)
Heart failure	488 (15.4%)	404 (45.3%)
Hypertension	958 (30.2%)	567 (63.6%)
Intracranial bleeding	9 (0.3%)	4 (0.4%)
Liver disease	60 (1.9%)	45 (5.1%)
Myocardial infarction	50 (1.6%)	79 (8.9%)
Peripheral arterial disease	42 (1.3%)	104 (11.7%)
Pulmonary embolism	139 (4.4%)	189 (21.2%)
Renal disease	303 (9.6%)	292 (32.8%)
Stroke/TIA	95 (3.0%)	106 (11.9%)
Valvular heart disease	2340 (73.9%)	339 (38.0%)
Other medication use (n, %)		
ACE inhibitor	154 (4.9%)	243 (27.3%)
ADP receptor inhibitor	140 (4.4%)	99 (11.1%)
ARB	715 (22.6%)	68 (7.6%)
Aldosterone antagonist	712 (22.5%)	40 (4.5%)
Allopurinol	65 (2.1%)	45 (5.1%)
Amiodarone	237 (7.5%)	174 (19.5%)
Aspirin	323 (10.2%)	632 (70.9%)
Beta-blocker	667 (21.1%)	655 (73.5%)
Calcium channel blocker, dihydropyridine	781 (24.7%)	86 (9.7%)
Calcium channel blocker, non-dihydropyridine	224 (7.1%)	20 (2.2%)
Diuretic, loop	1,625 (51.3%)	545 (61.2%)
Diuretic, thiazide	121 (3.8%)	34 (3..8%)
Insulin	61 (1.9%)	111 (12.5%)
Metformin	96 (3.0%)	19 (2.1%)
Nitrate	948 (29.9%)	68 (7.6%)
Other lipid-lowering medication	186 (5.9%)	34 (3.8%)
Statin	1,207 (38.1%)	453 (50.8%)
Sulfonylurea	178 (5.6%)	21 (2.4%)

Table 3. Characteristics of participants. The categorical variables, such as sex, comorbidities, and other medication use variables, are presented as numbers and percentages of patients with a specific sex, diagnosis, and medication, respectively. The remaining continuous variables are presented as the mean and standard deviation.

ADP, adenosine diphosphate; ARB, angiotensin receptor blocker; SD, standard deviation; TIA, transient ischemic attack.

3.2 Model Performance

The MAE and accuracy at the threshold for both datasets are listed in **Table 4**. The following MAEs were calculated for the internal validation set: XGBoost, 0.9; random forest, 1.0; Artificial neural nets (ANN), 1.0; and linear regression, 1.0. All models had better prediction performance than baseline (MAE of 1.3). Using the external validation dataset, the following MAEs were achieved: baseline, 1.8; random forest, 1.8; linear regression, 1.8; ANN, 2.0; and XGBoost, 1.9. Consequently, internal validation of the internal validation set from the AMC EMRs confirmed that all predictions of the artificial intelligence models had lower errors and higher accuracy than those made by physicians regarding MAEs and accuracy. However, the baseline showed similar or superior performance when compared with all machine learning models in terms of the MAE, in external validation derived from the MIMIC-III. The MAE box plots of the internal validation set and external validation set are shown in **Figure 3**.

	Internal validation set (n=634)				**External validation set (n=891)**			
	MAE	**Accuracy (e=0.5 mg)**	**Accuracy (e=1.0 mg)**	**Accuracy (e=1.5 mg)**	**MAE**	**Accuracy (e=0.5 mg)**	**Accuracy (e=1.0 mg)**	**Accuracy (e=1.5 mg)**
XGBoost	0.9	51.3	72.2	83.8	1.9	33.4	48.3	57.7
Random forest	1.0	49.5	68.9	81.5	1.8	35.6	44.6	56.8
Linear regression	1.0	48.7	70.0	84.9	2.0	31.9	45.3	58.1
Neural net	1.0	46.5	69.2	85.0	1.8	29.7	45.3	60.0
Physicians	1.3	32.2	57.3	69.4	1.8	37.1	48.3	53.5

Table 4. Model performance according to the MAE and accuracy. We conducted model performance evaluations of the internal validation set and external validation using the MAE and calculated the model prediction accuracy using three thresholds (0.5 mg, 1.0 mg, 1.5 mg). MAE, mean absolute error.

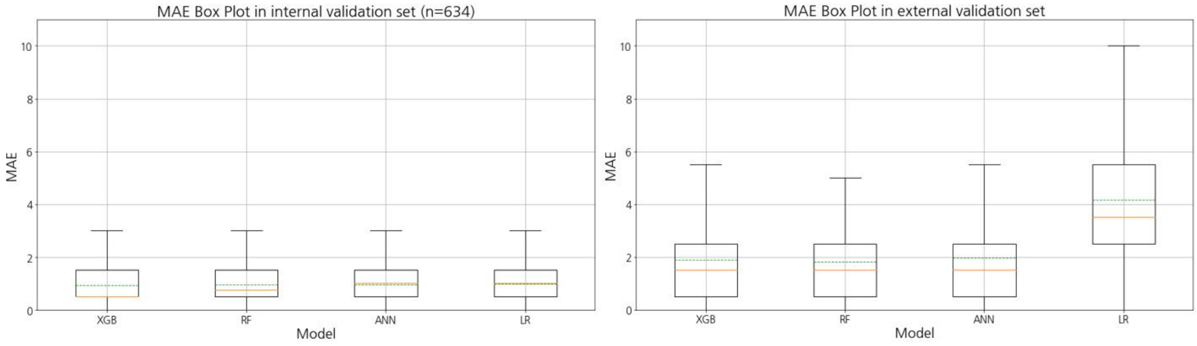

Figure 3. Box plot using the MAE. Performance abilities of the models based on the MAE were visualized using the internal validation (n=634) and external validation (n=891) sets. The top line represents the maximum value. The middle line represents the third quartile. The orange line represents the median. The bottom line represents the first quartile. The green dotted line represents the mean value. MAE, mean absolute error.

3.3 Model Interpretations

We examined the SHAP values of the 20 features with the most impact on model predictions using the beeswarm plot (**Fig. 4**). Additionally, we confirmed interactions between each feature using weight, height, and sex, because they are primary factors used to determine the warfarin dosage (**Fig 5**). We also investigated the impact of features on individual predictions. We randomly selected 4 data points from the internal validation set with no missing values, and our models accurately predicted all of them. Subsequently, we calculated the SHAP values using a waterfall plot to explain individual predictions. The waterfall plot explains the influence of each feature on individual predictions (**Fig. 6**). The patient of **Fig. 6a** was a 64-year-old male diagnosed with atrial fibrillation who received angiotensin receptor blockers, amiodarone, beta-blockers, diuretics (loop), and statins. His height was 171.8 cm. His weight was 89.5 kg. His heart rate, oxygen saturation, and respiration rate were 132 beats/minute, 99%, and 16 breaths/minute, respectively. The patients of **Fig. 6b** was a 52-years-old female diagnosed with valvular heart disease. Her height was 154cm and weight was 54.5kg. Her systolic blood pressure and diastolic blood pressure were 118 mmHg, 78 mmHg, respectively. The patient of **Fig. 6c** was a 72-years-old female diagnosed with atrial fibrillation, hypertension, pulmonary embolism. She was prescribed the various of medications, including statin, ARB, calcium channel blocker and nitrate, diuretic (loop), calcium channel blocker. Her height was 157.8 cm and weight was 65.9 kg. The patient of **Fig. 6d** was a 73-year-old male diagnosed with heart failure and renal disease who received aldosterone antagonists and diuretics (loop). His height was 166 cm. His weight was 76 kg. His oxygen saturation, systolic blood pressure, and diastolic blood pressure were 100%, 134 mmHg, and 57 mmHg, respectively.

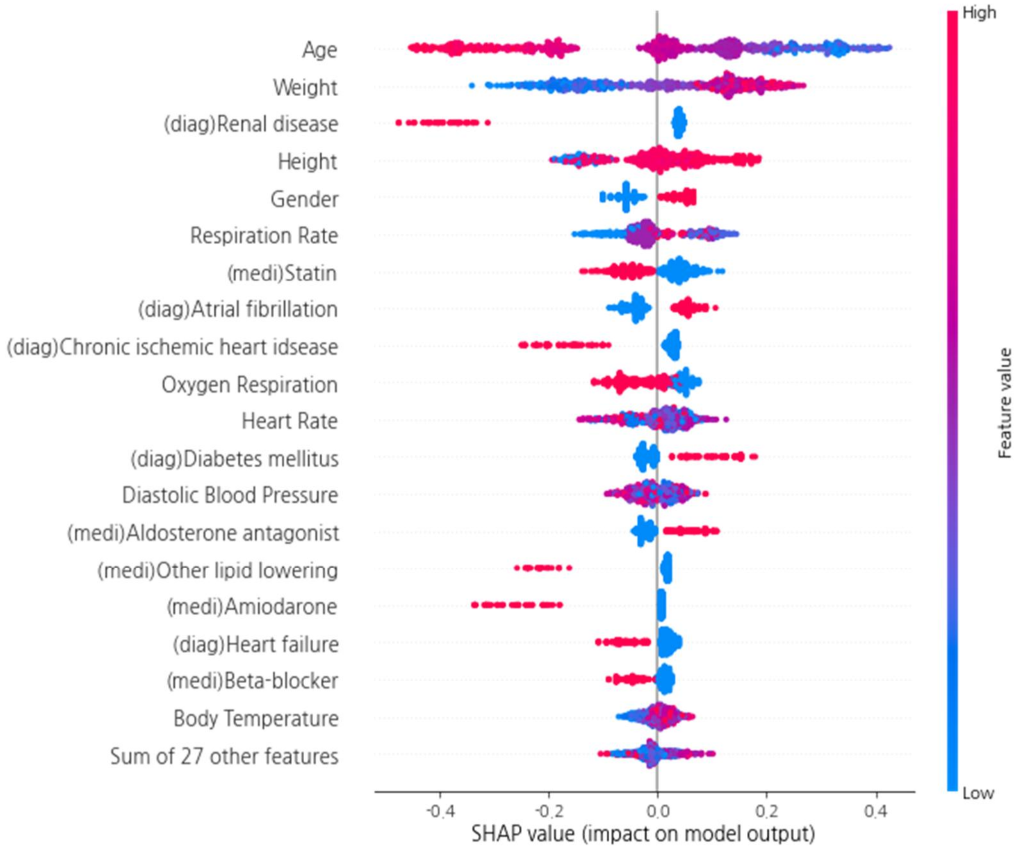

Figure 4. SHAP beeswarm plot of the 20 main features affecting the predictions of the XGBoost model. Features are ranked in descending order based on the absolute value of their influence on the XGBoost model. The x-axis indicates SHAP values. Each dot denotes a data point. Colors represent high values (red) or low values (blue) of specific data points. SHAP, Shapley additive explanations.

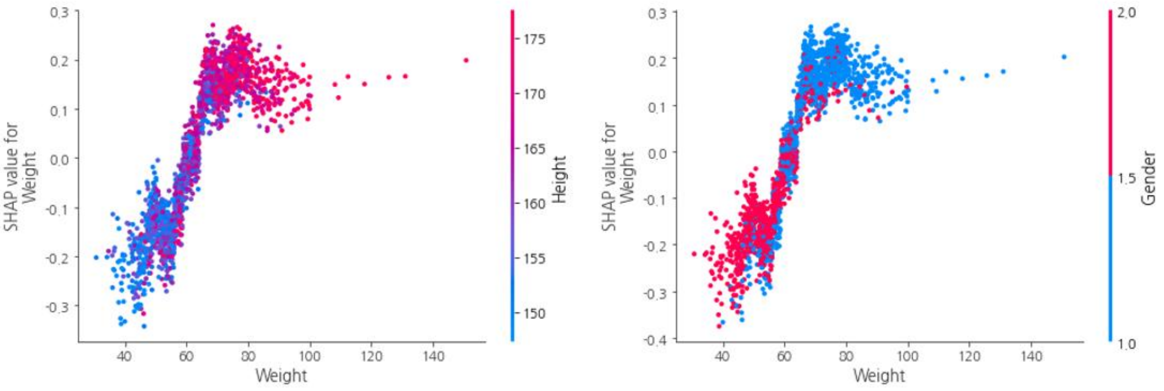

Figure 5. SHAP dependence plot. Each dot represents a data point. The x-axis indicates each feature. The y-axis indicates the SHAP value of each feature. Colors indicate whether a feature interacting with another feature had a high value (red) or low value (blue). SHAP, Shapley additive explanations.

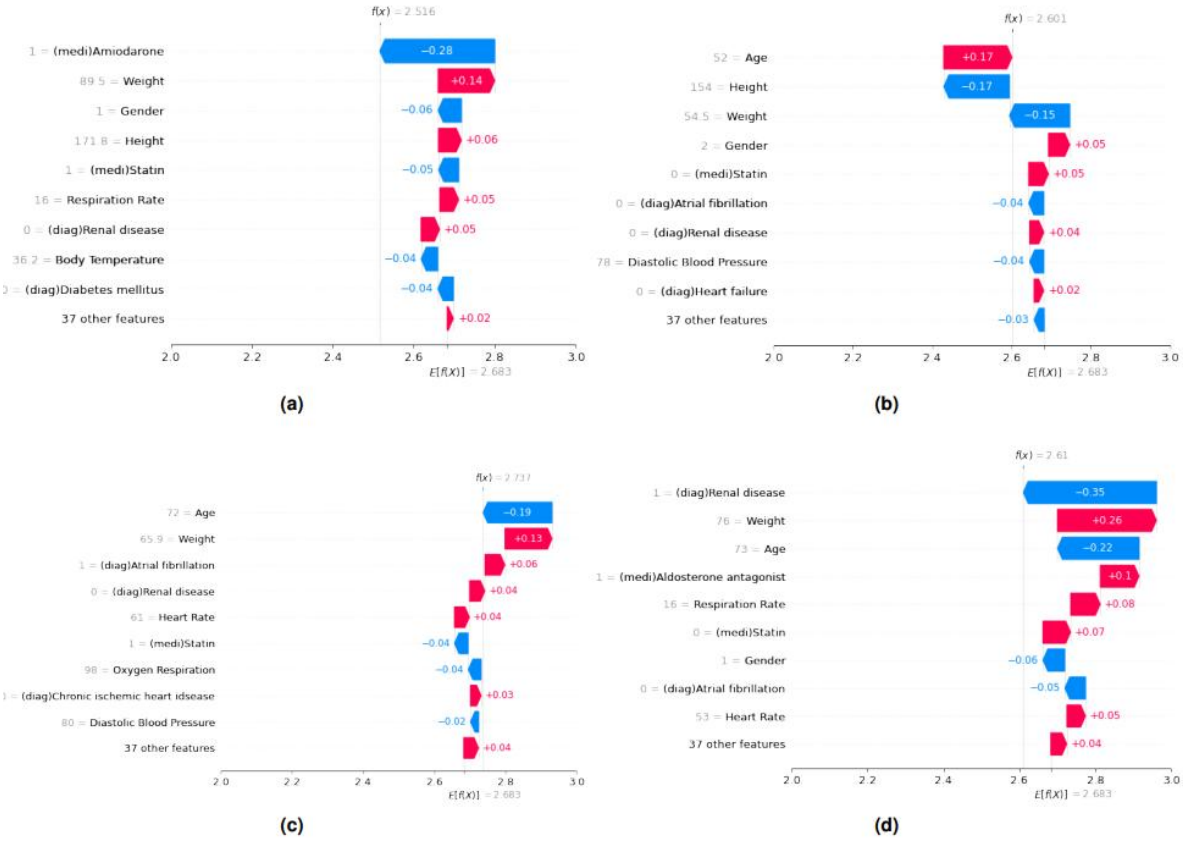

Figure 6. SHAP waterfall plot. The x-axis represents the individual warfarin dosage prediction of the models. The y-axis represents the input features of the models. E[f(x)] (2.571) represents the baseline value, which is the model output of the entire dataset, and f(x) represents the individual model output for each patient. Each arrow indicates whether a specific feature increased (red) or decreased (blue) the warfarin dosage. SHAP, Shapley additive explanations.

3.4 Comparison of Models' and Physicians' Predictions

Finally, we collected the predictions of the model and those of the five physicians for 40 data points (**Fig. 7**). First, intraclass correlation coefficient (ICC) of the physicians' predictions was calculated to measure the interrater agreement (**Table 5**). ICC is a value between 0 and 1, where values below 0.5 indicate poor reliability, between 0.5 and 0.75 moderate reliability, between 0.75 and 0.9 good reliability, and any value above 0.9 indicates excellent reliability [17]. As a result, ICC value of 0.16 was obtained and the consistency about warfarin dosage decision of the physicians was poor. Consequently, physicians tend to focus on a specific warfarin dosage range based on their clinical experience, thus leading to significant variability in the distribution of dosage predicted by physicians. Next, prediction accuracy of the models and the five physicians was compared. Physicians accurately predicted approximately 10 of 40 samples, achieving 25% accuracy; however, the models demonstrated 50% accuracy. This indicated that it can be difficult for physicians to determine the appropriate warfarin discharge dosage based on the 46 clinical features obtained within 2 days of hospitalization.

	Intraclass correlation coefficient (95% CI)	p-value
Five physicians' prediction	0.16 (0.02-0.35)	2.3e-17

Table 5. ICC of the predictions by the physicians for 40 data points.

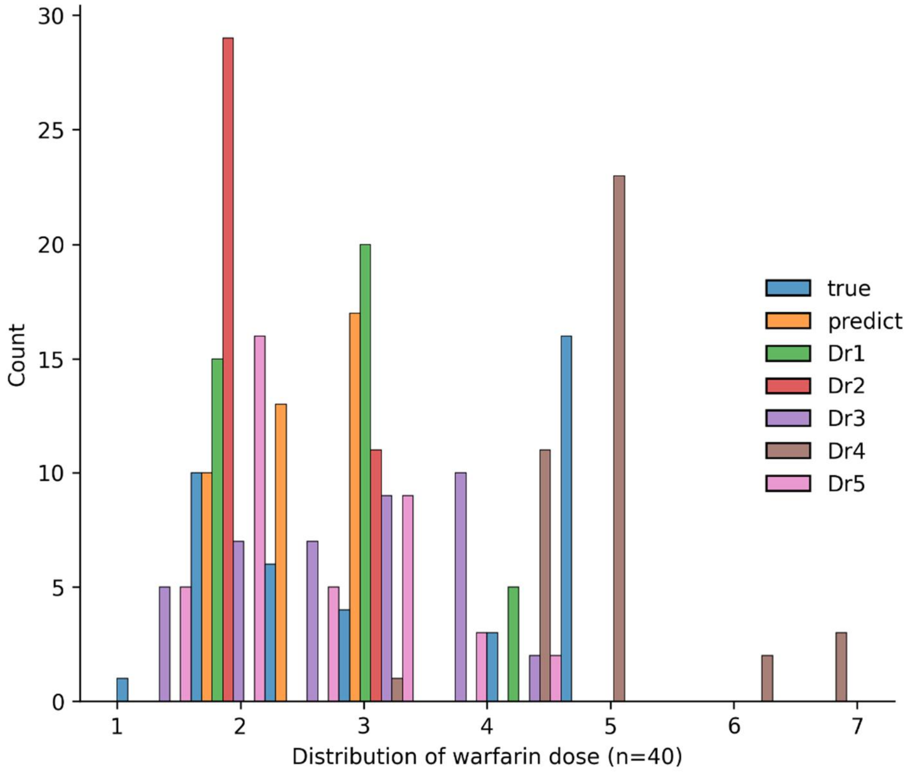

Figure 7. Distribution of warfarin dosage predicted by the physicians and models. We identified the predictions of five physicians and that of our model using 40 data points. The actual warfarin dosage at discharge (blue) was distributed evenly within 1-6mg. The average prediction dosage of our model (yellow) was 2·6 mg, and the maximum prediction dosage of our model was 3 mg; furthermore, it had 50% accuracy when predicting dosage of 2-3 mg. The predictions of the five physicians were diverse, and their accuracy was approximately 25%.

4. Discussion

It is important to determine the appropriate warfarin dosage for each individual to maximize its efficacy and safety [44]. However, the time required to determine the warfarin dosage varies for each person because of the influence of individual genetic and clinical factors [45]. In particular, VKORC1, CYP2C9, and CYP4F2 polymorphisms account for approximately 40% of the variability in warfarin responses, and new genetic variants that have not yet been identified are assumed to account for approximately 50% of the variability; however, clinical factors that can be considered in clinical practice affect the response to warfarin by only approximately 10% [46]. Nonetheless, genetic testing is costly and time-consuming; it is not routinely performed to determine the warfarin dosage [47]. Additionally, ICC value we analyzed in this study showed that there is inconsistency about warfarin dosage decided by each physician. Indeed, physicians rely on INR measurements and their medical experience to determine the appropriate warfarin dosage; they do not consider the genetic characteristics of patients rigorously [48]. Consequently, this study is significant because the warfarin discharge dosage was rapidly determined using only clinical information obtained within two days of hospitalization, unlike previous pharmacogenetic models. Additionally, our models showed similar or superior performance when compared to other warfarin dosage models (**Appendix Table 4, Data 1**). It demonstrated that the models can make appropriate warfarin dosing decisions without the same level of effort as physicians who would consider various factors such as the INR value. These results are likely attributable to the successful selection of important variables able to interact with warfarin from our initial clinical data obtained through discussions with experienced clinical experts and effective utilization of refined variables.

However, we were not able to successfully perform external validation in the MIMIC-III. The random forest and linear regression models had equal model performance with baseline (1.8) and the XGBoost and artificial neural network showed poor performance than baseline. Although we anticipated these issues, we had to conduct external validation using the MIMIC-III because it is the only public database available in clinical settings. Because the two datasets were created in different countries, there may have been differences in clinical information, such as various races. Our models trained using data of an East Asian population could not consider differences in races and physical characteristics in the MIMIC-III. Furthermore, there may have been differences in drug information because we categorized drug ingredients according to Korean standards, which may not reflect all drugs with the same efficacy used in the United States. Finally, because the MIMIC-III is an ICU database, warfarin dosage at discharge for general ward patients and those for ICU patients are not the same. Physicians who work in the MIMIC-III might already know the INR values of patients measured in general ward. The differences in the datasets make it difficult to directly compare the predictions made by physicians using the two datasets. Consequently, the models' weights that trained in training set were not applicable for external validation set. We need to conduct further study that train the models after standardizing a multi-center data to improve models' generalizability. In further work, we have to obtain a multi-center cohort and standardize a common cardiovascular registry to conduct an external validation successfully.

In conclusion, we developed 3 machine learning models and 1 deep learning model using data of a model development cohort from a tertiary hospital in Korea to predict warfarin discharge dosage. Although successful external validation was not performed during this study because our models could not consider differences in both of the clinical settings, our models showed outstanding predictive ability and outperformed physicians regarding accuracy using the internal validation set of the same institution as the training set. In internal validation set with MAE, XGBoost models achieved 0.9, and random forest and linear regression, artificial neural network models achieved 1.0, whereas physicians achieved 1.3. As a result, all of our models outperformed physicians. Therefore, our models could alleviate the difficulties encountered by physicians when determining the warfarin discharge dosage of patients admitted to the hospital for the first time. Especially, it might be effective tools that help physicians choose more accurate and personalized warfarin doses for an East Asian population while reducing unnecessary treatment durations and preventing warfarin overdose and adverse drug events.

Chapter 2.

Database integrated between EMR (Electronics Medical Record) and Open Targets Platform for pharmacogenomic study

Abstract

Background: To reduce the cost and time involved in the traditional drug development process, drug development research that uses artificial intelligence (AI) is actively being conducted. Besides open-source databases, electronic medical records (EMRs, which serve as a warehouse for medical data generated in clinical settings) can provide real-world evidence and can be used in the process of drug discovery. In this study, we constructed a new platform by linking information from the Open Targets Platform, a well-known open-source medicine database, with EMRs at Asan Medical Center, Seoul, Republic of Korea.

Objective: The study objective was to contribute to personalized drug development, enabling future research on drug–protein interactions; therefore, we established a clinical database that could be used as an AI drug development platform.

Methods: The 'Target-Disease evidence' cancer dataset was selected from the Open Targets Platform. Then, clinical data for patients who had been prescribed at least one of the 110 drugs in the Open Targets Platform were extracted from the ABLE system at the Asan Medical Center. Finally, the two datasets were integrated, based on drug information.

Results: The study population comprised 1,380 participants, and the database contained 53 drugs associated with 91 targets and 60 cancer types. The database also contained 10 types of genetic variation: loss-of-function mutation, missense mutation, amplification, frameshift deletion, frameshift insertion, nonsense mutation, splice-site mutation, in-frame deletion, in-frame insertion, and nonstop mutation.

Conclusions: Data obtained from the open-source database were consistent with data obtained from the clinical setting; however, there were some cases where information was found only in the clinical setting and not in the open-source database. Therefore, a database with the integration of open-source and clinical information can provide new evidence that can be used in future AI drug research.

Keywords: EMR; electronic medical record; heterogeneous database; Open Targets Platform;

1. Introduction

Drug development is an interdisciplinary field that requires the integration and coordination of biology, chemistry, and pharmacology [49]. Typically, drug development involves selecting disease targets (eg, target discovery), target validation, lead compound identification, lead compound optimization, preclinical development, clinical trials, and registration [40]. Target discovery explains the process of discovering compounds that show efficacy against clinically-validated targets [51]. Through lead compound optimization, approximately five out of 1,000 compound candidates proceed to preclinical stages [52]. Absorption, distribution, metabolism, excretion, and toxicity studies are conducted through laboratory and animal experiments as a preclinical process [53]. Clinical trials, including Phase 1, 2, and 3 trials, are conducted in humans. Registration refers to Food and Drug Administration (FDA) review and approval [53]. Thus, the overall drug development process takes an average of 10–15 years, requires millions to billions of dollars in investment, and has a success rate of only 1 in 5,000 to 1 in 10,000 [54].

To expedite and improve the efficiency of drug development, artificial intelligence (AI) drug discovery, which uses clinical data and suitable AI algorithms to develop drugs, is now being applied [55, 56]. AI drug discovery is a very useful method that can produce results quickly, and with high accuracy and low cost, compared to traditional drug-discovery methods [57]. AI can search for an almost infinite number of substances at the candidate substance search stage, which can significantly reduce costs and time [58]. Additionally, AI can help identify patterns and interrelationships inherent in data, thereby opening up new possibilities for detecting disease targets and for drug redesign that would be difficult to uncover using conventional methods [59]. Because of these advantages, AI drug development is attracting attention as a promising method for faster and more efficient drug development [60].

AI requires big data for learning, such that data reliability is crucial [61]. Therefore, it is important to refine a large amount of prior knowledge-based data for the intended purpose [62]. By using refined big data in AI, a rapid drug-development method can be proposed, and the drug-development period can be significantly shortened [63]. Among medical big data, electronic health records (EHRs) provide vast amounts of patient data, such as prescription results, diagnostic codes, prescriptions, and physician opinions [64]. As EHR databases become more standardized and integrated across multiple hospital systems, they are gaining attention as data sources to be analyzed when evaluating patient treatment and building early disease-prediction models [65]. Therefore, we built a new database, optimized for AI drug development and based on clinical practice, by linking existing public databases related to pharmaceutical information and EMRs from Asan Medical Center (AMC), Seoul, Republic of Korea; our database has the potential to be used in future AI drug development research (**Fig. 1**).

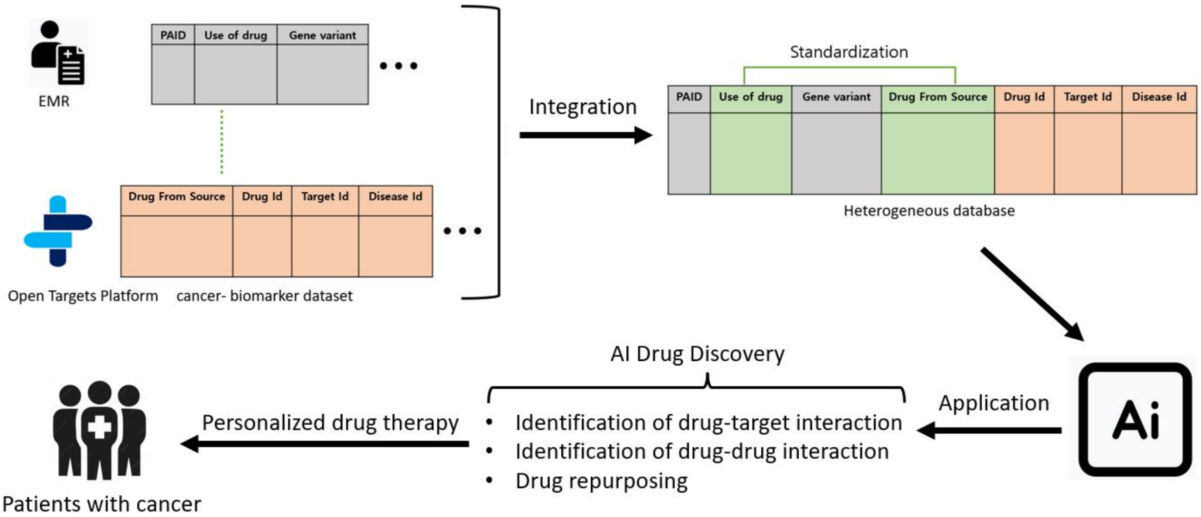

Figure 1. Study overview. In this study, we integrated Open Targets Platform and EMR derived from Asan Medical Center. In Open Targets Platform, cancer-biomarker dataset included 110 FDA approved anticancer drugs

was used. The database was mapped based on drugs, because drug ingredients from EMR and 'Drug From Source' from Open Targets Platform were equivalent. ChEMBL drug Id, Ensembl gene Id, Experimental Factor Ontology disease Id were used key identifier of drug, target, disease, respectively in Open Targets Platform. Finally, a heterogeneous database was generated and could be added information of clinical association using key identifiers. This heterogeneous database could be useful in future AI research for drug development, such as identification of drug-target interaction or drug-drug interaction, drug repurposing. EMR, electronic medical record; AI, artificial intelligence;

2. Methods

2.1. Data source: Open Targets Platform

Open Targets Platform is a large public database of pharmacologic knowledge that includes a variety of drug, target, and disease-related information, and details about biologic pathways, and gene ontology [66]. Due to its vast collection of data, the Open Targets Platform can be used as a tool to discover new drug targets and indications and can be integrated with other databases. The 'Target-Disease evidence' cancer dataset of the Open Targets Platform contains 110 drugs approved by the FDA for the treatment of various cancers and comprises a two-dimensional dataset using a many-to-many approach, with each drug having one or more targets. In summary, drug–target disease association data in the platform include 110 drugs, 182 targets, and 76 disease-information presents. ChEMBL identification (ID), Ensembl gene ID, and Experimental Factor Ontology disease ID were used for drug ID, target ID, and disease ID, respectively, in the Open Targets Platform. Consequently, this dataset was selected as the main dataset of the Open Targets Platform for our research because we could use both drug ID and target ID as the primary identifier and could merge drug information based on the identifier.

2.2 Cohort extraction: AMC EMRs

We extracted clinical and genetic data from EMRs at AMC. The AMC dataset comprised patients aged >19 years and admitted to the cardiovascular or thoracic and cardiovascular surgery departments of AMC between January 1, 2018 and October 31, 2020. In clinical practice, many drugs are used for purposes other than their originally approved indications [67]. Therefore, we excluded drugs in the EMRs that were not listed in the Open Targets Platform, and we decided to use only EMR data, including cancer biomarkers, for the 110 FDA-approved drugs in the 'Target Disease evidence' cancer dataset. Ultimately, we collected a final study population of 1,380 individuals by including only patients with genetic testing records and diagnosed with a specific cancer. We constructed a single dataset for these patients that included: diagnostic and medication information, and genetic testing results. This study obtained approval from the ARC institutional review board (approval no. 2021-0321). All experiments were performed in accordance with relevant guidelines and regulations.

2.3 Integration between EMR and open targets platform

Because drug ingredient name in the EMR matched the 'drugFromSource' variable in the Open Targets Platform, the two datasets were integrated based on drug type. Drugs in the EMR that had no information in the 'Target-Disease evidence' cancer dataset were excluded, so that only FDA-approved drugs with an association between drug and target were included. Thus, linked 73 drug information in the AMC dataset among 110 drugs. Despite the availability of EMRs, patients without a genetic-test result or specific cancer diagnosis were excluded. Finally, the integrated database of information from EMRs and the Open Targets Platform comprised 53 drugs associated with 91 targets and 60 cancer types. The final dataset comprised 1,262,425 rows and 20 columns.

3. Results

A new drug information database was created that could be used for AI-based drug development research, and that was constructed by integrating the AMC EMR database with the Open Targets Platform, which contains information on 110 drugs that can treat cancer. We linked this data with the genetic-test results for 1,380 patients who received 57 of these 110 drugs at AMC. As each drug could have one or more target genes or target cancers, the information included: 91 genes previously identified as targets for the drugs, and 60 types of cancer that the drugs could treat. The mapping ratios for integration of the Open Targets Platform with the EMR database are shown in **Table 1**.

Variables	Total codes (N)	Mapped code (N)	Mapping ratio (%)
Medication	110	57	52
Diagnosis	80	60	75
Target	182	91	50

Table 1. Mapping ratios for the Open Targets Platform. Information on FDA-approved drugs available in both the Open Targets Platform and AMC EMRs was used. Among the 110 drugs in the Open Targets Platform: over 80 were FDA-approved for cancer treatment and 182 biomarker targets were clearly identified; drug information for 73 drugs could be used for mapping with AMC EMR data, but patients without genetic-testing data were excluded, resulting in the final extraction of drug information for 53 drugs. The total number of targets identified to interact with the selected 53 drugs was 91, and there were 60 types of target cancer. AMC, Asan Medical Center; EMR, electronic medical record; FDA, Food and Drug Administration.

There were 533 gene mutations (not necessarily associated with drugs or cancer) and 39 cancer types obtained from the EMR database. **Table 2** lists the six most frequently prescribed drugs, and the six most frequently diagnosed cancer types, in the clinical setting.

Drug	Count (%)	Diagnosis	Count (%)
Cisplatin	361 (26.2)	Non-small cell lung cancer	323 (23.4)
Paclitaxel	326 (23.6)	Gastric cancer	213 (15.4)
Octreotide	192 (13.9)	Ovarian cancer	194 (14.1)
Gemcitabine	160 (11.6)	Pancreatic cancer	137 (9.9)
Bevacizumab	155 (11.2)	Malignant brain tumor	95 (6.9)
Capecitabine	134 (9.7)	Other	68 (4.9)

Table 2. Frequency of drug and diagnosis derived from EMR database (N=1,380). It showed that the count and percentage of the top 6 drugs and cancer types that were most frequently prescribed and diagnosed, respectively, among 1380 participants in the EMR database. EMR, electronic medical record;

To examine the characteristics of patients who received each drug, the six most frequently used drugs were selected, and the most frequently diagnosed cancer types within these six drug categories were identified (**Fig. 2**).

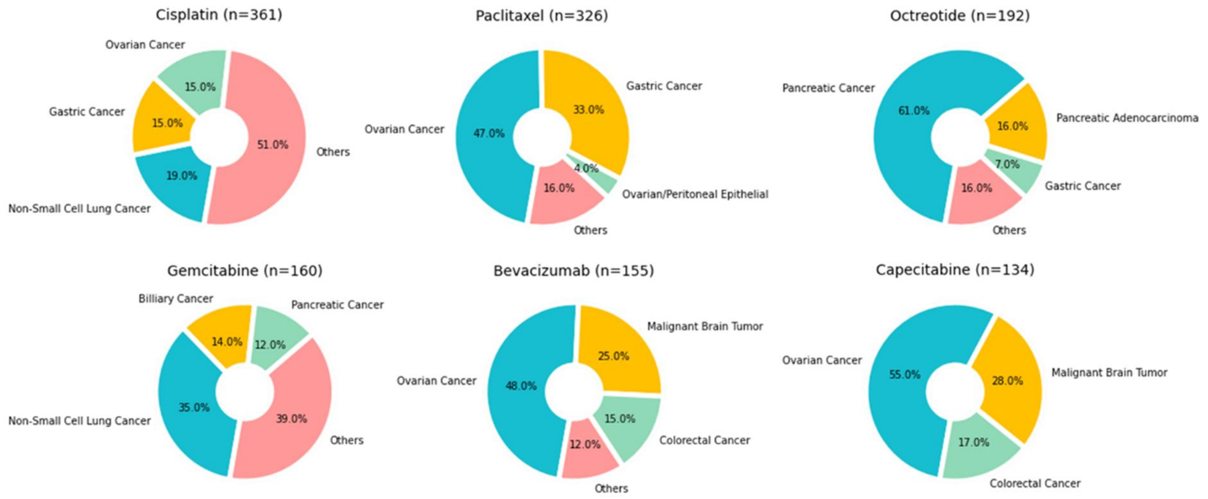

Figure 2. Cancer types, shown as the percentage of patients receiving each of the six most frequently prescribed drugs. n represents the number of patients receiving each drug.

Data for 10 types of genetic variation were extracted from the EMRs: loss-of-function mutation, missense mutation, amplification, frameshift deletion, frameshift insertion, nonsense mutation, splice-site mutation, in-frame deletion, in-frame insertion, and nonstop mutation. The top 10 types of genetic variation, and the top 10 specific gene mutations, are listed in **Table 3**.

Variant type	**Count (%)**	**Gene variant**	**Count (%)**
Missense mutation	1,369 (99.2)	TP53	906 (65.7)
Amplification	855 (62.0)	CDKN2A	316 (22.9)
Loss-of-function mutation	614 (44.5)	KRAS	306 (22.2)
Nonsense mutation	514 (37.2)	EGFR	286 (20.7)
Frameshift deletion	491 (35.6)	LRP1B	252 (18.3)
Splice-site mutation	330 (23.9)	BRCA2	232 (16.8)
In-frame deletion	281 (20.4)	NOTCH3	228 (16.5)
Frameshift insertion	230 (16.7)	CDKN2B	227 (16.4)
In-frame insertion	74 (5.4)	MYC	224 (16.2)
Nonstop mutation	4 (0.3)	PIK3CA	222 (16.1)

Table 3. The 10 most frequently observed types of genetic variation and specific gene mutations (N=1,380). The most common structural genetic variations were single-nucleotide variant missense mutations, and the most common gene mutation was TP53 mutation. Loss, loss-of-function mutations;

In addition, we examined the frequency of gene mutations in patients treated with the top six prescribed drugs (**Fig. 3**). Almost all the 1,380 participants had missense mutations, and more than half of the participants had amplification and TP53 mutations.

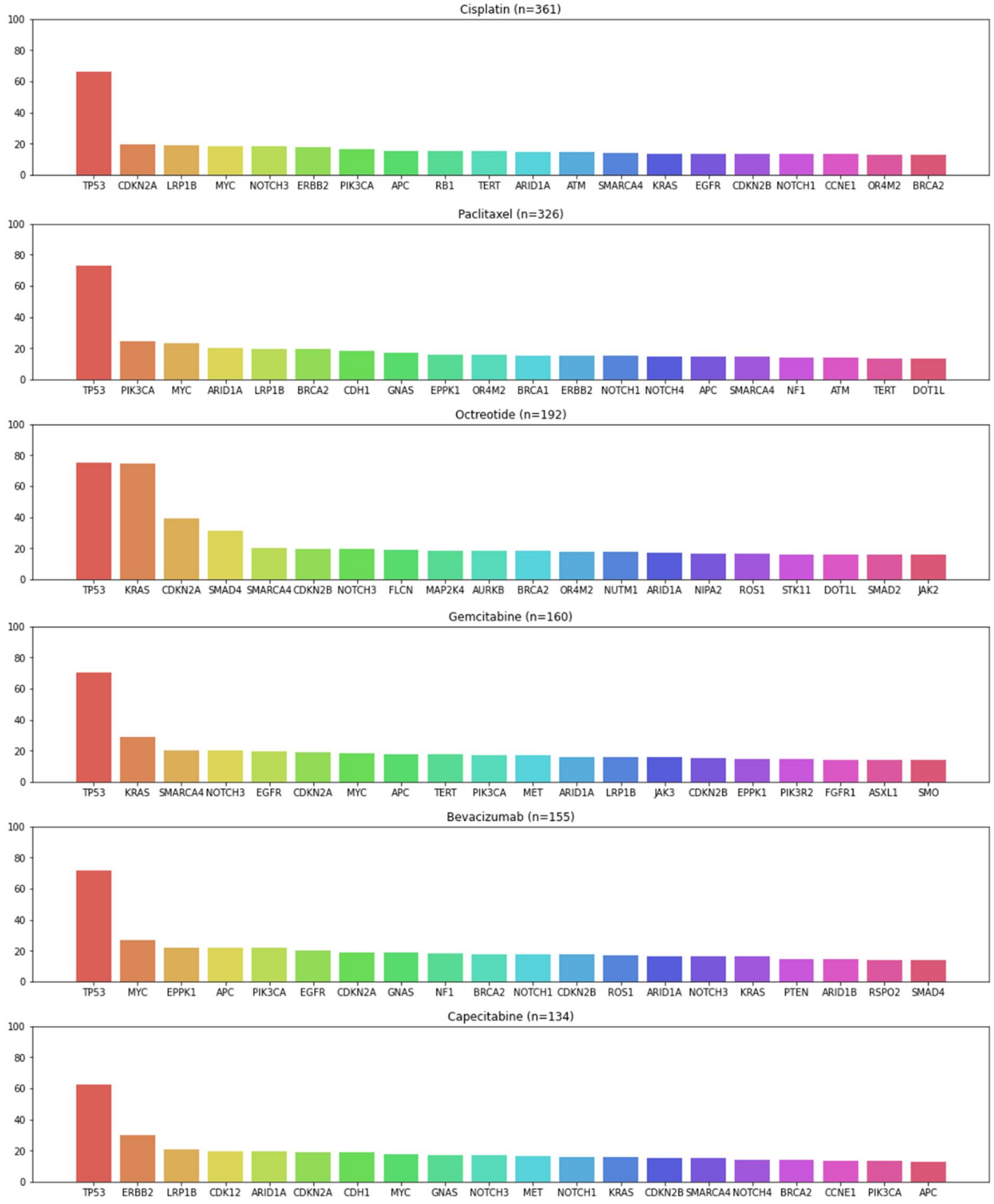

Figure 3. Frequency of specific gene mutations in patients receiving each of the top six prescribed drugs. Data derived from the electronic medical record database; n represents the number of patients receiving each drug.

4. Discussion

If EMR or open-source data alone were used in AI drug development, important information might be lost. However, the heterogeneous database constructed in this study has significant potential to improve the quality of big data and use additional information in future AI drug development research. If EMR data alone were available, paclitaxel would be used as a drug targeting TUBB3 to treat bladder cancer, but since TUBB3 mutation testing is not conducted at AMC, EMR data alone could miss information about the known interaction between drug and target [68]. Conversely, if using only the Open Targets Platform, clinically important information on CDKN2A that could affect the prescribing of anticancer drugs would be unavailable and might result in missing information on drug targets [69]. We decreased information loss by integrating the ACR EMR and Open Targets Platform databases and found agreement between the two databases. Consequently, the novel, heterogeneous database could potentially be used to provide real-world evidence.

In our overall data population, 65% of people had a TP53 mutation. TP53, a tumor suppressor gene, is the most mutated gene in cancer and can progress specific types of cancer when a mutation occurs, thus making cancer treatment more difficult [70]. Therefore, TP53 has been studied as a target for various anticancer agents and, in this study, the correlation between TP53 and cancer was clinically confirmed once again.

Cisplatin has therapeutic effects on various cancers, such as bladder, breast, lung, and ovarian, and targets genes such as ERCC2, FANCC, BRCA1, ERCC4, MDM2, ATM, BRCA2, RB1, and TP53 [71-72]. Same as this, in clinical practice, some individuals who received cisplatin were found to have the corresponding cancer and gene variant. Paclitaxel is renowned as an effective medication for bladder carcinoma [73], but in clinical practice in this study, it was used to treat various cancers, such as ovarian, gastric, lung, and kidney. Octreotide is renowned as a medication that can treat meningioma by targeting the NF2 gene [74-75]. However, in the clinical setting in this study, patients diagnosed with ovarian or pancreatic cancer and with an NF2 mutation received octreotide (ie, without a diagnosis of meningioma). Gemcitabine targets KRAS, TP53, and SLC29A1 and has therapeutic effects against pancreatic and bladder carcinoma [76-77]. Same as this, in the clinical field, patients diagnosed with pancreatic cancer were prescribed this medication. Bevacizumab is approved to treat malignant peripheral nerve sheath tumor, glioma, and schwannoma in patients with VHL, NF1, and NF2 mutations [78-81], but in the clinical setting in our study, bevacizumab was used in patients with renal carcinoma and VHL, NF1, and NF2 mutations. Capecitabine targets DPYD [82], and our study confirmed that capecitabine was prescribed to individuals with DPYD mutations in the clinical field. Thus, not only can actual research results be confirmed by clinical data, but it is also expected that new clinical evidence can be discovered by using additional genetic and disease information from the study participants.

Importantly, using our database in AI research for new drug development is expected to enable personalized drug treatment for patients based on their disease and genetic information. Also, our database may facilitate the discovery of new indications for currently used drugs, and the prediction of new drug–target interactions (DTIs) or drug–drug interactions. Indeed, in a future study, we will use our database to predict new DTIs through the development of AI models.

Conclusion

EMR has a significant potential to contribute to personalized drug treatment due to include extensive medical big data, such as diagnosis code, medication information, genotype data, physician's opinion. Additionally, AI can understand complicated big data and find inherent patterns and correlations in big data. Finally, we decided that collaborate AI with EMR to contribute to personalized drug treatment.

This study aimed to realize personalized drug treatment using electronic medical record (EMR) data and developed two tools for that purpose. The first study focused on the early prediction of discharge dosage for the anticoagulant warfarin using machine learning models. Four machine learning models suitable for dosage prediction were developed, and through internal validation, it was confirmed that the model predictions were more accurate than those of actual clinical experts. Additionally, the SHAP (SHapley Additive exPlanations) technique was employed to analyze the key variables influencing the model predictions, providing insights into the decision-making process similar to that of healthcare professionals. Besides, when the same data was presented to individual physicians, significant variations in dosage decisions based on their individual medical experiences were observed. Furthermore, model showed that its prediction accuracy was twice as high as that of physicians, when used internal validation set. Therefore, the clinical utility of the models was demonstrated in East Asian population. Additionally, we have to conduct a further study that make our models consider not only East Asian, but also other races.

The second study focused on the integration of EMR and pharmaceutical information databases to construct a novel clinical field-based heterogeneous database. FDA-approved anticancer agents and associated target gene information from the Open Targets Platform database was used. We standardized drug elements between the EMR and Open Targets Platform, the two databases were linked based on drug information. Finally, we extracted the correlations between 57 anticancer agents, 60 cancer types, and 91 genetic mutations. Moreover, we contained additional information, such as diagnostic information and genetic test results of actual patients receiving anticancer drugs. In conclusion, a platform was built that could utilize both clinical and genetic characteristics of patients in real-world clinical settings for AI drug development. In further work, we will perform AI drug development, such as drug repurposing or drug-target interaction identification using the novel database.

We have concluded that this study can improve patient outcomes, minimize adverse drug reactions, and optimize drug therapies, ultimately leading to better overall healthcare delivery and patient care.

Reference

1. Pirmohamed, M. Warfarin: almost 60 years old and still causing problems. Br. journal clinical pharmacology 62, 509 (2006).

2. Pirmohamed, M., Kamali, F., Daly, A. K. & Wadelius, M. Oral anticoagulation: a critique of recent advances and controversies. Trends pharmacological sciences 36, 153–163 (2015).

3. Gage, B. F., Fihn, S. D. & White, R. H. Management and dosing of warfarin therapy. The Am. journal medicine 109, 481–488 (2000).

4. Glurich, I., Burmester, J. K. & Caldwell, M. D. Understanding the pharmacogenetic approach to warfarin dosing. Hear. failure reviews 15, 239–248 (2010).

5. Gage, B. et al. Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. Clin. Pharmacol. & Ther. 84, 326–331 (2008).

6. Pavani, A. et al. Artificial neural network-based pharmacogenomic algorithm for warfarin dose optimization. Pharmacogenomics 17, 121–131 (2016).

7. Roche-Lima, A. et al. Machine learning algorithm for predicting warfarin dose in caribbean hispanics using pharmacogenetic data. Front. pharmacology 10, 1550 (2020).

8. Tong, H. Y. et al. A new pharmacogenetic algorithm to predict the most appropriate dosage of acenocoumarol for stable anticoagulation in a mixed spanish population. PloS one 11, e0150456 (2016).

9. Grossi, E. et al. Prediction of optimal warfarin maintenance dose using advanced artificial neural networks. Pharmacogenomics 15, 29–37 (2014).

10. Saleh, M. I. & Alzubiedi, S. Dosage individualization of warfarin using artificial neural networks. Mol. Diagn. & Ther. 18, 371–379 (2014).

11. Hernandez, W. et al. Ethnicity-specific pharmacogenetics: the case of warfarin in african americans. The pharmacogenomics journal 14, 223–228 (2014).

12. Alzubiedi, S. & Saleh, M. I. Pharmacogenetic-guided warfarin dosing algorithm in african-americans. J. cardiovascular pharmacology 67, 86–92 (2016).

13. Martes-Martinez, C. et al. Cost-utility study of warfarin genotyping in the vachs affiliated anticoagulation clinic of puerto rico. Puerto Rico health sciences journal 36, 165–172 (2017).

14. Hu, Y.-H., Wu, F., Lo, C.-L. & Tai, C.-T. Predicting warfarin dosage from clinical data: A supervised learning approach. Artif. intelligence medicine 56, 27–34 (2012).

15. Johnson, A. et al. Mimic-iii, a freely accessible critical care database sci. Data 3, 10–1038 (2016).

16. Willmott, C. J. & Matsuura, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Clim. research 30, 79–82 (2005).

17. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. chiropractic medicine 15, 155–163 (2016).

18. Avecilla, S. T. Transfusion management of patients receiving antithrombotic therapy. In Transfusion Medicine and Hemostasis, 343–349 (Elsevier, 2019).

19. Stage, T. B., Brøsen, K. & Christensen, M. M. H. A comprehensive review of drug–drug interactions with metformin. Clin. pharmacokinetics 54, 811–824 (2015).

20. Dunvald, A.-C. D. et al. Initiation of glucose-lowering drugs reduces the anticoagulant effect of warfarin–but not through altered drug metabolism in patients with type 2 diabetes. medRxiv 2022–10 (2022).

21. Romley, J. A. et al. Association between use of warfarin with common sulfonylureas and serious hypoglycemic events: retrospective cohort analysis. Bmj 351 (2015).

22. Dimakos, J. et al. Concomitant use of sulfonylureas and warfarin and the risk of severe hypoglycemia: Population-based cohort study. Diabetes Care 45, e131–e133 (2022).

23. Wang, M. et al. Drug–drug interactions with warfarin: A systematic review and meta-analysis. Br. journal clinical pharmacology 87, 4051–4100 (2021).

24. Kean, M., Krueger, K., Parkhurst, B., Berg, R. & Griesbach, S. Assessment of potential drug interactions that may increase the risk of major bleeding events in patients on warfarin maintenance therapy. J Pharm Soc Wis 21, 44–8 (2018).

25. McDonald, M., Au, N., Wittkowsky, A. & Rettie, A. Warfarin–amiodarone drug–drug interactions: determination of [i] u/ki, u for amiodarone and its plasma metabolites. Clin. Pharmacol. & Ther. 91, 709–717 (2012).

26. Marengoni, A. et al. Understanding adverse drug reactions in older adults through drug–drug interactions. Eur. J. Intern. Medicine 25, 843–846 (2014).

27. Nutescu, E., Chuatrisorn, I. & Hellenbart, E. Drug and dietary interactions of warfarin and novel oral anticoagulants: an update. J. thrombosis thrombolysis 31, 326–343 (2011).

28. Greenblatt, D. J. & von Moltke, L. L. Interaction of warfarin with drugs, natural substances, and foods. The J. Clin. Pharmacol. 45, 127–132 (2005).

29. Limdi, N. A. et al. Warfarin dosing in patients with impaired kidney function. Am. J. Kidney Dis. 56, 823–831 (2010).

30. Shah, S. et al. Comparative effectiveness of direct oral anticoagulants and warfarin in patients with cancer and atrial fibrillation. Blood advances 2, 200–209 (2018).

31. Gulløv, A. L., Koefoed, B. G. & Petersen, P. Bleeding during warfarin and aspirin therapy in patients with atrial fibrillation: the afasak 2 study. Arch. internal medicine 159, 1322–1328 (1999).

32. Snipelisky, D. & Kusumoto, F. Current strategies to minimize the bleeding risk of warfarin. J. blood medicine 89–99 (2013).

33. Venables, W. N., Ripley, B. D., Venables, W. & Ripley, B. Tree-based methods. Mod. applied statistics with S-Plus 303–327 (1999).

34. Patro, S. & Sahu, K. K. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462 (2015).

35. Jain, A. K., Mao, J. & Mohiuddin, K. M. Artificial neural networks: A tutorial. Computer 29, 31–44 (1996).

36. Yan, X. & Su, X. Linear regression analysis: theory and computing (world scientific, 2009).

37. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794 (2016).

38. Breiman, L. Random forests. Mach. learning 45, 5–32 (2001).

39. LaValle, S. M., Branicky, M. S. & Lindemann, S. R. On the relationship between classical grid search and probabilistic roadmaps. The Int. J. Robotics Res. 23, 673–692 (2004).

40. James, G., Witten, D., Hastie, T. & Tibshirani, R. An introduction to statistical learning, vol. 112 (Springer, 2013).

41. Peng, C.-Y. J., Lee, K. L. & Ingersoll, G. M. An introduction to logistic regression analysis and reporting. The journal educational research 96, 3–14 (2002).

42. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 (2018).

43. Arrieta, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Inf. fusion 58, 82–115 (2020).

44. Ravvaz, K., Weissert, J. A., Ruff, C. T., Chi, C.-L. & Tonellato, P. J. Personalized anticoagulation: optimizing warfarin management using genetics and simulated clinical trials. Circ. Cardiovasc. Genet. 10, e001804 (2017).

45. Jonas, D. E. & McLeod, H. L. Genetic and clinical factors relating to warfarin dosing. Trends pharmacological sciences 30, 375–386 (2009).

46. Li, X. et al. Precision dosing of warfarin: open questions and strategies. The pharmacogenomics journal 19, 219–229 (2019).

47. Bussey, H. I., Wittkowsky, A. K., Hylek, E. M. & Walker, M. B. Genetic testing for warfarin dosing? not yet ready for prime time (2008).

48. Kuruvilla, M. & Gurk-Turner, C. A review of warfarin dosing and monitoring. Bayl. Univ. Med. Cent. Proc. 14, 305–306 (2001).

49. Mohs, R. C. & Greig, N. H. Drug discovery and development: Role of basic biological research. Alzheimer’s & Dementia: Transl. Res. & Clin. Interv. 3, 651–657 (2017).

50. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. Br. journal pharmacology 162, 1239–1249 (2011).

51. Emmerich, C. H. et al. Improving target assessment in biomedical research: the got-it recommendations. Nat. reviews Drug discovery 20, 64–81 (2021).

52. Jorgensen, W. L. The many roles of computation in drug discovery. Science 303, 1813–1818 (2004).

53. Wu, F. et al. Computational approaches in preclinical studies on drug discovery and development. Front. chemistry 8, 726 (2020).

54. Deore, A. B., Dhumane, J. R., Wagh, R. & Sonawane, R. The stages of drug discovery and development process. Asian J. Pharm. Res. Dev. 7, 62–67 (2019).

55. Deng, J., Yang, Z., Ojima, I., Samaras, D. & Wang, F. Artificial intelligence in drug discovery: applications and techniques. Briefings Bioinforma. 23, bbab430 (2022).

56. Paul, D. et al. Artificial intelligence in drug discovery and development. Drug discovery today 26, 80 (2021).

57. Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. Expert. opinion on drug discovery 16, 949–959 (2021).

58. Chen, W., Liu, X., Zhang, S. & Chen, S. Artificial intelligence for drug discovery: Resources, methods, and applications. Mol. Ther. Acids (2023).

59. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. Drug discovery today 23, 1241–1250 (2018).

60. Hemati, S., Masoumi, M., Masoumi, A. & Zareh, R. The application of artificial intelligence to simulate and model health and medical systems: A technical survey.

61. Daniel, E. O'leary. Technol. for Knowl. Assim. Marshall Sch. Business, Univ. South. California, Los Angeles 90089–0441 (2000).

62. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. From data mining to knowledge discovery in databases. AI magazine 17, 37–37 (1996).

63. Masuda, Y. et al. Adaptive enterprise architecture for the digital healthcare industry: A digital platform for drug development. Information 12, 67 (2021).

64. Jamoom, E., Heisey-Grove, D., Yang, N. & Scanlon, P. Physician opinions about ehr use by ehr experience and by whether the practice had optimized its ehr use. J. health & medical informatics 7 (2016).

65. Yao, L., Zhang, Y., Li, Y., Sanseau, P. & Agarwal, P. Electronic health records: Implications for drug discovery. Drug discovery today 16, 594–599 (2011).

66. Ochoa, D. et al. Open targets platform: supporting systematic drug–target identification and prioritisation. Nucleic acids research 49, D1302–D1310 (2021).

67. Young-Mee Lee, Ji-Hae Lee and Hyeon S. Son. (2016). Bioinformatics analysis for drug repositioning. The Korean Journal of Public Health, 53(2), 19-27.

68. Miyata, Y. et al. Expression of class iii beta-tubulin predicts prognosis in patients with cisplatin-resistant bladder cancer receiving paclitaxel-based second-line chemotherapy. Anticancer. research 38, 1629–1635 (2018).

69. Lin, J.-C., Liu, T.-P. & Yang, P.-M. Cdkn2a-inactivated pancreatic ductal adenocarcinoma exhibits therapeutic sensitivity to paclitaxel: a bioinformatics study. J. Clin. Medicine 9, 4019 (2020).

70. Olivier, M., Hollstein, M. & Hainaut, P. Tp53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb. perspectives biology 2, a001008 (2010).

71. Michalska, M., Schultze-Seemann, S., Kuckuck, I., Katzenwadel, A. & Wolf, P. Impact of methadone on cisplatin treatment of bladder cancer cells. Anticancer. Res. 38, 1369–1375 (2018).

72. Florea, A.-M. & Büsselberg, D. Cisplatin as an anti-tumor drug: cellular mechanisms of activity, drug resistance and induced side effects. Cancers 3, 1351–1371 (2011).

73. Sideris, S. et al. Efficacy of weekly paclitaxel treatment as a single agent chemotherapy following first-line cisplatin treatment in urothelial bladder cancer. Mol. Clin. Oncol. 4, 1063–1067 (2016).

74. Graillon, T. et al. Octreotide therapy in meningiomas: in vitro study, clinical correlation, and literature review. J. neurosurgery 127, 660–669 (2016).

75. Johnson, D. R. et al. Phase ii study of subcutaneous octreotide in adults with recurrent or progressive meningioma and meningeal hemangiopericytoma. Neuro-oncology 13, 530–535 (2011).

76. Amrutkar, M. & Gladhaug, I. P. Pancreatic cancer chemoresistance to gemcitabine. Cancers 9, 157 (2017).

77. Waters, A. M. & Der, C. J. Kras: the critical driver and therapeutic target for pancreatic cancer. Cold Spring Harb. perspectives medicine a031435 (2017).

78. Martin, E., Flucke, U. E., Coert, J. H. & van Noesel, M. M. Treatment of malignant peripheral nerve sheath tumors in pediatric nf1 disease. Child's Nerv. Syst. 36, 2453–2462 (2020).

79. Morris, K. A. et al. Toxicity profile of bevacizumab in the uk neurofibromatosis type 2 cohort. J. neuro-oncology 131, 117–124 (2017).

80. Tao, J., Sun, D., Dong, L., Zhu, H. & Hou, H. Advancement in research and therapy of nf1 mutant malignant tumors. Cancer cell international 20, 1–8 (2020).

81. Hrisomalos, F. N., Maturi, R. K. & Pata, V. Long-term use of intravitreal bevacizumab (avastin) for the treatment of von hippel-lindau associated retinal hemangioblastomas. The open ophthalmology journal 4, 66 (2010).

82. Dean, L. & Kane, M. Capecitabine therapy and dpyd genotype. (2020).

Appendix

Table 1. Features used for model training and validation

Sex
Age
(diag)Angina
(diag)Arrhythmia
(diag)Atrial fibrillation
(diag)Cancer
(diag)Chronic ischemic heart disease
(diag)Chronic lung disease
(diag)Diabetes mellitus
(diag)Dyslipidemia
(diag)Heart failure
(diag)Hypertension
(diag)Intracranial bleeding
(diag)Liver disease
(diag)Myocardial infarction
(diag)Peripheral arterial disease
(diag)Pulmonary embolism
(diag)Renal disease
(diag)Stroke / TIA
(diag)Valvular heart disease
(medi)ACE inhibitor
(medi)ADP receptor inhibitor
(medi)ARB
(medi)Aldosterone antagonist
(medi)Allopurinol
(medi)Amiodarone
(medi)Aspirin
(medi)Beta-blocker
(medi)Calcium channel blocker, dihydropyridine
(medi)Calcium channel blocker, non-dihydropyridine
(medi)Diuretic, loop
(medi)Diuretic, thiazide
(medi)Insulin
(medi)Metformin
(medi)Nitrate
(medi)Other lipid lowering
(medi)Statin
(medi)Sulfonylurea
Height
Weight
Heart Rate
Oxygen Respiration
Systolic Blood Pressure
Diastolic Blood Pressure
Body Temperature
Respiration Rate

Table 2. ICD-10 diagnostic codes used to categorize diseases.

Diagnosis	**ICD-10 code**
Myocardial infarction	I21, I22, I23
Hypertension	I10, I11, I12, I13, I15
Diabetes mellitus	E10, E11, E12, E13, E14
Dyslipidemia	E78
Chronic ischemic heart disease	I25
Angina	I20
Heart failure	I42, I43, I50
Valvular heart disease	I05, I06, I07, I08, I09, I34, I35, I36, I37, I38, I39
Peripheral arterial disease	I70, I73, I74, I77, I79
Stroke / TIA	I63, I64, I65, I66, I67, I68, G45, G46, H34
Intracranial bleeding	I60, I61, I62
Atrial fibrillation	I48
Arrhythmia	I44, I45, I47, I49
Liver disease	K70, K71, K72, K73, K74, K75, K76, K77
Renal disease	N03, N04, N05, N10, N11, N12, N13, N14, N15, N16, N17, N18, N19, Z49, Z940, Z992
Chronic lung disease	J40, J41, J42, J43, J44, J45, J46, J47, J60, J61, J62, J63, J64, J65, J66, J67, J84
Cancer	C
Pulmonary embolism	I26, I27

Table 1. Information of the four models hyperparameter.

Model	**Hyper-parameter**
XGBoost	Learning rate: 0.4 Max depth: 3 Objective function: regression with squared log loss Minimum sum of instance weight needed in a child: 10 L1 regularization term on weights: 0.6 Minimum loss reduction required to make a further partition on a leaf node of the tree: 0.6
Random Forest	Number of estimators: 250 Max depth: 7 Max leaf nodes: 9 Minimum of leaf: 7 Number of samples to draw from dataset to train each base estimator: 0.6 Minimum of impurity decrease to split the nodes: 0.0005 Number of minimum samples to split nodes: 4 Minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node: 0.4
Artificial neural network	Activation function: identity function Hidden layer: (300, 300, 300, 200, 200) Learning rate: 0.01 Solver: Minibatch gradient descent Batch size: 32 Max iteration: 200
Linear Regression	Intercept: used

Appendix Table 4. Comparison with prior works. Our models were compared with previous models that developed using both of genetic and clinical data or only clinical data. Performance metric was used MAE. The information of study population' race was included. The number of data participants included the number of model development and validation cohort. * represents the best performance model. ANN: artificial neural network;

	Number of data instances	Features	Study population	algorithm	MAE
Choi et al*	3,168	Clinical	South Korean	XGBoost	0.9
Choi et al	3,168	Clinical	South Korean	ANN	1.0
Choi et al	3,168	Clinical	South Korean	Random Forest	1.0
Choi et al	3,168	Clinical	South Korean	Linear Regression	1.0
Gage et al [5]	1,307	Clinical + pharmacogenetic	Caucasian, African American	Regression	1.0
Gage et al [5]	1,307	Clinical	Caucasian, African American	Regression	1.5
Pavani1[6]	240	Clinical + pharmacogenetic	Indian	ANN	1.97
Roche-Lima et al [7]	190	Clinical + pharmacogenetic	Caribbean Hispanics	Random forest	4.7
Tong et al [8]	685	Clinical + pharmacogenetic	Spanish	Multiple linear regression	3.5
Tong et al [8]	685	Clinical	Spanish	Multiple linear regression	5.0
Grossi et al [9]	377	Clinical + pharmacogenetic	Caucasian	ANN	5.72
Saleh et al [10]	4,271	Clinical + pharmacogenetic	Multi-ethnic	ANN	9.0
Hernandez et al [11]	349	Clinical + pharmacogenetic	African-American	Multivariate regression	10.9
Alzubiedi et al [12]	163	Clinical + pharmacogenetic	African-American	Linear regression	10.8
Alzubiedi et al [12]	163	Clinical + pharmacogenetic	African-American	ANN	10.9

Appendix Data 1. Comparison with prior works

We identified previous regression models that predict warfarin dosage as numerical target and use MAE as a performance metrics, to compare accurately between our models and other models. Gage et al developed a multiple regression model that predict warfarin dosage in derivation cohort (N=1,015), included Caucasian and African American, Hispanic, and validated the models in validation cohort (N=292). Both of pharmacogenetic and clinical model were developed and reached a MAE of 1.0, 1.5, respectively. Pavani et al developed an artificial neural network using ten genetic variables as inputs and therapeutic warfarin dosage as the output in Indian population (N=240). Roche-Lima et al collected cardiovascular patients of 190 Caribbean Hispanic were >21 years old and developed seven machine learning algorithms that predict warfarin dosing in Caribbean Hispanics using pharmacogenetic data. Among them, random forest regressor (RFR) significantly outperformed all other models with a MAE of 4.74. Tong et al recruited 685 patients who diseased atrial fibrillation or thromboembolic venous disease in a Spanish population using the data last 3 consecutive months and used multiple linear regression. Both of pharmacogenetic and clinical model were developed and internally validated with a MAE of 3.5, 5.0, respectively in a validation cohort (N=129). Grossi et al collected 377 patients who were over 18 years old and treated with warfarin in Caucasian population and developed an artificial neural network to predict a optimal warfarin maintenance dose. The final model reached a MAE of 5.72. Saleh recruited 4,271 multi-races patients who received warfarin and developed an artificial neural network using both of genotyping and clinical data. The artificial neural network model reached a MAE of 9.0. Hernandez et al generated pharmacogenomic warfarin dosing model using clinical and genotyping retrospective data from a derivation cohort of 349 African Americans patients were >= 18 years and the model reached with a MAE of 10.9mg/week. Alzubiedi et al collected demographic, clinical, and genetic data from 163 African-American patients with a stable warfarin dose. They developed both of a multiple linear regression model and artificial neural network model with MAE of 10.8, 10.9 respectively. Whereas, our XGBoost model achieved 0.9 with MAE and outperformed aforementioned algorithms. Additionally, our models provided more appropriate warfarin dosage than those initially prescribed by physicians using clinical data within 2 days of hospitalization.

국문 요약

많은 환자들이 질병치료를 위해 약물치료를 병행한다. 그러나 각 개인이 가지고 있는 고유의 특징이 모두 다르기 때문에 동일한 약물로 치료를 받더라도, 약물 반응이 모두 다르게 나타난다. 이로 인해, 환자마다 약물의 용법을 달리함으로써, 개인의 특성을 고려하는 맞춤형 약물치료가 필요하다. 맞춤형 약물치료가 잘 이루어지면, 약물 부작용은 최소화하고 치료 효과는 최대화할 수 있다는 장점이 있다. 그러므로, 이미 사용되고 있는 약물 또는 새롭게 개발되어야 하는 약물은 모두 맞춤형 약물로 사용될 수 있어야 한다. 따라서, 본 논문에서는 병원의 전자의무기록 (EMR) 데이터를 활용하여 맞춤형 약물치료를 실현하기 위해 두 가지 연구를 수행하였다.

첫 번째 연구에서는, 항응고제 warfarin 의 퇴원용량을 조기 예측하는 기계학습 모델을 개발하고 검증하였다. 먼저 약물 용량에 적절한 네 가지 기계학습 모델을 개발하였으며, 내부검증을 통해 모델 예측이 실제 임상의의 예측보다 정확도가 높은 것을 확인하였다. 또한 SHAP (SHapley Additive exPlanations)기법을 사용하여 모델 예측에 영향을 주는 주요 변수를 분석함으로써 모델 예측을 설명하였다. 이를 통해, 모델의 용량결정 과정이 실제 의사와 매우 유사하다는 것을 확인하였다. 마지막으로, 동일한 데이터를 의사에게 제시했을 때 각 의사의 개별적 의료 경험에 따라 용량 결정에 변동성이 큰 것을 확인하였으며, 모델의 예측 정확도가 의사보다 2 배 더 높다는 임상적 유용성을 입증하였다.

두 번째 연구에서는, EMR 과 의약정보 데이터베이스를 통합함으로써 새로운 임상현장 기반의 database 를 구축하였다. Open Targets Platform 이라는 데이터베이스에서 FDA 승인을 받은 항암제 및 관련 표적 유전자 정보를 추출하였다. 또한 EMR 과 Open Targets Platform 의 약물 성분 표준화를 통해, 두 데이터베이스를 약물 기준으로 연계하였다. 그 결과, 57 개의 항암제와 관련된 60 개의 암 종류 및 91 개의 유전자 돌연변이의 연관성을 확인할 수 있었다. 또한 실제 항암제 처방환자들의 또 다른 진단정보 및 유전자검사 결과를 확인함으로써, 실제 임상현장에서 환자들의 임상적, 유전적 특징을 모두 활용할 수 있는 플랫폼을 구축하였다.

본 연구를 통해, 전자의무기록 데이터를 활용하여 맞춤형 약물치료를 도울 수 있는 tool 을 두 가지 개발하였다. 최적 warfarin 용량을 예측하는 인공지능 모델은 임상의의사결정보조시스템으로 사용됨으로써, 불필요한 치료기간을 감소시키고 약물부작용을 사전에 방지하는 데 기여할 수 있다. 또한 전자의무기록 데이터와 의약정보 데이터베이스를 통합하여 새롭게 구축된 heterogeneous database 는 인공지능 신약개발에 활용될 수 있을 것으로 기대된다.