Machine Learning for Improved Warfarin Dosing in South Africa

Gianluca Truda
TRDGIA001
University of Cape Town
trdgia001@myuct.ac.za

ABSTRACT

The anticoagulant drug warfarin is an effective preventative treatment for strokes and heart attacks. However, it requires individualised dosing that accounts for numerous factors. This is typically performed by human experts, but software dosing is more effective and resource efficient. This study evaluated the accuracy of 17 learning algorithms on both a South African warfarin dataset and the international IWPC dataset. The first 10 algorithms used default or manually-optimised hyperparameters, but the remaining 7 algorithms were developed using genetic programming. These automated algorithms produced the most accurate models and outperformed the best published results in this field. This study also examined the effects of parameter sets and missing data treatments on model accuracy, which informed guidelines on how to implement dosing models in a South African clinical context.

KEYWORDS

Supervised learning, pharmacogenetics, IWPC, PathCare, automated dosing, autoML, TPOT, computational medicine.

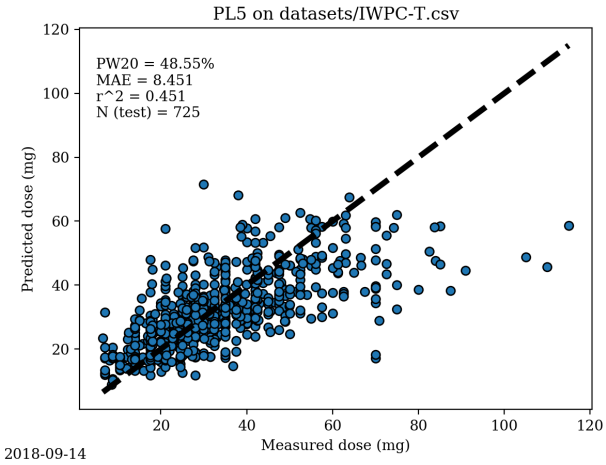

Figure 1: Example of the predictive performance of the autoML-generated model that set new performance benchmarks (PW20 of 48.81% and MAE of 8.41) on the international IWPC dataset.

1 INTRODUCTION

Many individuals suffer from obstructive blood clots that lead to strokes and heart attacks. The standard method for treating these conditions is the use of anticoagulant drugs, such as warfarin. Whilst effective, the drug has a narrow therapeutic range and severe side-effects at extreme concentrations. This makes the precise dosing of warfarin an important concern for clinicians. Unfortunately, warfarin metabolism differs across individuals based on age, weight, genetics, diet, drug interactions, and various pre-existing conditions [18, 47]. International standards and dosing protocols have attempted to formalise the dosing procedure, and software tools exist to assist clinicians in making informed decisions, but the high individual variability and risk of severe bleeding make the development of more accurate dosing methods an ongoing priority. Many studies have looked at applying statistical models to the problem of individualised warfarin dosing. Accurate models would improve the ability of clinicians to prescribe the correct warfarin doses to their patients, whilst minimising the time required to do so. They would also reduce the risk of severe haemorrhaging in patients and the number of visits required to establish a therapeutic dose. This is especially relevant in South Africa, where haematopathologists and resources are in short supply. Unfortunately, warfarin datasets are small and noisy, which requires the use of specialised data transformations and highly-optimised learning algorithms. Breakthrough techniques are of significance to the medical research community and could lead to a future in which warfarin therapy is safer and more automated – freeing up valuable human resources to address other medical challenges.

This study investigated avenues for improved warfarin dosing in South African patients, focusing on three research aims: (1) To evaluate if models performed as well as, or better than, human experts on a dataset of South African patients. (2) To determine which data-manipulation practices caused models to produce the most accurate and robust warfarin dose predictions in general. And (3) to evaluate if models trained using contemporary techniques could outperform the best published results in warfarin dosing[1].

A comprehensive review of the 16 most notable studies in automated warfarin dosing revealed promising algorithms and strategies that were utilised in investigating the research aims. Using a standard international warfarin dataset [48], alongside warfarin records provided by the pathology group PathCare, this study evaluated 17 promising machine learning algorithms.

Section 2 provides background knowledge on warfarin metabolism and current dosing practices before examining promising avenues for both manual and automated machine learning. Section 3 outlines the methods used to develop and evaluate learning algorithms on local and international datasets. In section 4, a novel machine learning pipeline is presented. This, along with manually-optimised algorithms, is evaluated on multiple datasets in section 5. Finally, the paper highlights key concerns before drawing conclusions and outlining avenues for future work.

UCT, 2018, CSC4000W

[1] From Liu et al. 2015 [25]

2 BACKGROUND

Warfarin causes an anticoagulant effect by inhibiting vitamin K-dependent clotting factors. It has achieved popularity and widespread usage due to its superior bioavailability and relatively predictable onset, but has a very narrow therapeutic range and poses serious risk of haemorrhaging at high doses [17]. There are two different phases of warfarin dosing – initiation and maintenance. The initiation dose is the quantity of warfarin administered to a patient beginning anticoagulation therapy, whilst the maintenance dose is used to keep a patient in a therapeutic range once warfarin has already saturated their system. Simple algorithms guide clinicians in administering an appropriate initiation dose [2]. The maintenance dose, however, can be incredibly unpredictable. At this phase of treatment, the patient is tested very infrequently, so the maintenance dose must be precisely tuned to keep them within therapeutic levels. To complicate this, there is substantial variation in how individuals respond to warfarin. Many clinical factors, such as age, race, weight, height, and smoking status must be taken into account when determining warfarin dosage [17]. There are also genetic factors. Around 40% of the individual variation in dose requirement can be attributed to polymorphisms in only two genes – *CYP2C9* and *VKORC1* [18]. There are also 26 foods and drugs known to interact with warfarin. These include a number of antibiotics, cardiac drugs, three drugs that act on the central nervous system, vitamin K-rich foods, and even large quantities of avocado [47].

2.1 INR and Traditional Dosing

To standardise the process of anticoagulant monitoring, the World Health Organisation established the international normalised ratio (INR) as a universal reference value [22]. INR is regionally calibrated [33], so direct comparison of values across geography and demographics is possible. The recommended therapeutic range for oral anticoagulation therapy is an INR between 2.0 and 3.0 for most patients [1]. It has been found that using tighter target ranges for maintenance dosing does not achieve any improvement in anticoagulation control [28]. For the purposes of treatment, any measures within this ±0.5 range are equally good – an important factor to consider when evaluating the predictive performance of any dosing protocol.

Traditionally, maintenance doses are determined by clinicians on a individual basis and adjusted regularly to ensure an INR in the desired range [20]. Due to the low availability of haematopathologists in South Africa, organisations like PathCare send basic patient details (like age, sex, last INR, current medications) to a few remote experts who evaluate the case and send back a prescribed dosing regimen. The imprecision and error involved in manual dosing has resulted in a concerted effort to (at least partially) automate warfarin dosing [20].

2.2 Software-Assisted Dosing

There are two widely-used equations for warfarin dosing in clinical practice. Gage et al. [11] used an exponential function based on a dozen clinical and genetic factors; whilst the International Warfarin Pharmacogenetics Consortium (IWPC) [23] used least-squares linear regression to develop their equation based on age, height, weight, race, and VKORC1 and CYP2C9 genotypes. It is now very common for medical staff to make use of a dosing algorithm to guide their decisions, since they have been found to increase anticoagulant control [21]. As a natural progression, many also make use of software tools to guide this process. Software-assisted dosing has been found to be safer, more therapeutic, and more cost-effective than manual dosing [19, 20, 34]. Whilst these tools are invaluable to clinicians, there exists much room for improvement. Since 2008, machine learning approaches have been shown to offer improved results over the current dosing software [27].

2.3 Machine Learning Algorithms

The goal of automated dosing is to provide a model with a vector of patient parameters (age, height, weight, etc.) and have it predict the target value – in this case, weekly warfarin dose in mg. In order to develop such a model, learning algorithms are provided with a set of labelled examples called the *training* set. Each example features the parameters for a patient as well as the true weekly dose for that patient. The learning algorithm performs regression on this data, producing a model of the relationships between the parameters and the weekly dose. This model is then tested on unseen data, with its predicted doses being compared to the true doses. Any model that can successfully *predict* warfarin doses given to patients in the past can be used to *suggest* doses for new patients in the future. This means that real-world performance of models can be accurately estimated without the need for clinical experiments.

2.3.1 Studies on Warfarin Dosing with Machine Learning. The 16 notable studies in this field [4, 12, 13, 15, 16, 23–25, 27, 29, 36, 38–40, 46, 50] were grouped into two categories: (1) those who used bespoke datasets, and (2) those who used the IWPC dataset. It is difficult to replicate the findings of the 13 studies in category 1, so the value of their findings was down-weighted accordingly. The three studies in category 2 [23, 25, 38] made use of the public IWPC dataset [48] and thus provided better insight into the relative performance of candidate algorithms. Linear regression (LR), support vector regression (SVR), lasso regression (LAS), and regression trees (RT) were popular in the literature. More recently, artificial neural networks (ANNs) – specifically the simplest kind, known as multilayer perceptrons (MLPs) – have been used to improve predictive performance [15, 27, 50].

2.3.2 Algorithm Optimisation. The learning algorithm chosen has a notable impact on model accuracy and robustness. Moreover, the precise hyperparameters chosen for the algorithm drastically affect performance. Selecting the best learning algorithm for the data and then optimising its hyperparameters is a challenge whenever machine learning is applied. The most common approach is for an expert to repeatedly retrain with various learning algorithms and tweak their parameters until performance is maximised. Another approach is the use of automatic machine learning (autoML), which employs meta-algorithms to automate the task of optimisation. One promising approach to *autoML* is genetic programming.

2.3.3 Genetic Programming for Optimisation. Genetic programming (GP) emulates the process of natural selection to optimise computer programs. In this case, the programs are machine learning algorithms and data processors. Each is encoded as a *gene*,

with parameters randomly changing according to a defined *mutation rate*. As with biological evolution, successful genes propagate through the population and the most successful combinations of those genes seed the next generation [3]. This increases performance over time. By simulating many generations with the right mutation rate and population sizes, novel programs with high performance are produced. GP has been shown to develop intelligent systems in a number of mathematical and computational domains [8, 10, 14, 41] and is of extreme interest to computer science in general. Figure 2 depicts how a Tree-based Pipeline Optimization Tool (TPOT) automates the nebulous and tedious aspects of the machine learning process [30]. Data preprocessors and learning algorithms are used as the genetic elements in TPOT and combine to produce machine learning *pipelines*. These handle the data from feature extraction through to parameter optimisation [30]. The TPOT framework was developed atop the Distributed Evolutionary Algorithms in Python (DEAP) framework [9] and has been shown to produce a significant improvement over basic machine learning methods, with little input or prior knowledge from users [30].

2.3.4 Combining Models with Ensemble Methods. Ensemble methods combine the outputs from several models to obtain better overall performance. This has been found effective in a number of warfarin dosing studies [4, 15, 25]. Some ensemble methods such as *random forest regression* are specific to a class of algorithms – in this case, regression trees – whilst others like *voting* and *boosting* are meta-algorithms that work for any homogeneous or heterogeneous collection of models [7, 35].

2.4 Pharmacogenetics

Studies on warfarin dose prediction use one of two input classes to develop models – either only clinical data, or both clinical and genetic data. Common *clinical* factors include age, body mass, and height. *Pharmacogenetic* factors include the genotypes for *CYP2C9* and *VKORC1*, which affect warfarin metabolism [18]. Many studies report improved model performance when using both clinical and pharmacogenetic data [23, 29, 32, 37, 45]. Two recent studies evaluated genotype-guided (pharmacogenetic) dosing under clinical conditions, finding that it brought patients to their therapeutic range more quickly and safely [44] and was significantly more effective for long-term anticoagulation therapy [6]. Unfortunately, genetic testing is still very expensive, resource intensive, and unavailable in many parts of the world [15].

2.5 Performance Benchmark

A notable study is that of Liu et al. in 2015 [25], which compared the average performance of 9 learning algorithms on the International Warfarin Pharmacogenetics Consortium (IWPC) dataset [48]. They filtered-out patients missing height, weight, age, or genotype data, and patients not at a stable warfarin dose – leaving 4798 patients remaining. Liu et al. then obtained seven clinical and two pharmacogenetic co-variates with step-wise regression. These were used as input parameters[2] for the models. Using libraries in R, they implemented 9 algorithms with default parameters – linear regression

[2]See 3.1.2

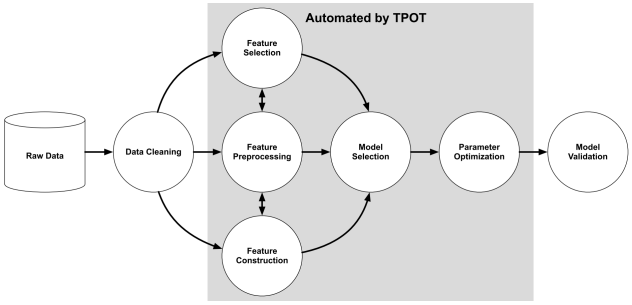

Figure 2: Illustration of the steps in the typical supervised learning workflow that are automated by the Tree-based Pipeline Optimization Tool (TPOT). Image from Olson et al. [30].

(LR), artificial neural network (ANN), regression tree (RT), multivariate adaptive regression splines (MARS), boosted regression tree (RT), support vector regression (SVR), random forest regression (RFR), lasso regression (LARS) and Bayesian additive regression tree (BART). They obtained the average performance of each algorithm[3] with 100 rounds of 80/20 re-sampling from the filtered dataset. Whilst the main focus of their study was evaluating a range of off-the-shelf algorithms across dosage ranges and racial groups, their top results[4] in the combined cohort served as a benchmark for performance on the IWPC data.

3 METHODOLOGY

This section details the process by which the three research aims were investigated, with experiments designed to falsify each hypothesis. It also describes the protocols used to develop, train, and evaluate machine learning systems in a manner that adhered to best practices.

3.1 Experiment Design

The study consisted of three distinct experimental lines, each with their own methodology and criteria.

3.1.1 Experiment 1: Humans vs. Algorithms. The first experiment evaluated the accuracy of experienced human clinicians compared to learned models. This was possible due to the fact that the PathCare dataset contained records of multiple visits for a small subset of patients (N = 1135). Those records were compared with the performance of models trained on a subset of the data (N = 3696). Because the human experts dosed the patients based only on their electronic records (with no direct contact), they had little informational advantage and their performance could be fairly compared with that of a model. In examples from other domains, learned models outperform experienced humans when datasets and feature-spaces are sufficiently large. It was, therefore, hypothesised that models derived from the dataset would fall at least within 5% of human performance. It is noted that this was not an audit of clinician performance in general, but rather relative performance given the state and quality of the provided dataset.

[3]In terms of PW20 and MAE (see 3.7).
[4]PW20 = 46.35%, MAE = 8.84.

3.1.2 Experiment 2: Data-Manipulations. This two-part experiment evaluated the effects of data-manipulations on the resulting models, investigating the use of different parameter sets as well as different treatments for missing data.

2A: Comparison of Models Across Parameter Sets. This experiment assessed the relationship between the feature-richness of a parameter set and the performance of the resulting model. This included three sets of parameters, which are illustrated in figure 3 and detailed below.

(1) **The set common to PathCare and IWPC (2)**: Age in years, sex
(2) **The easily-implemented clinical set (6)**: Age in years, sex, race, height in cm, weight in kg, smoking status
(3) **The set used by Liu et al. (9)**: Age in years, race, height in cm, weight in kg, smoking status, amiodarone use, VKORC1 genotype, CYP2C9 genotype, use of enzyme inducer[5]

Assuming that the features used were correlated to warfarin metabolism, it was predicted that there would be a positive relationship between the feature-richness of a parameter set and the performance of the resulting model.

2B: Comparison of Missing-Data Treatments. This experiment compared the effects of three different missing data treatments on model performance. The treatments were:

(1) **Excision**: Removing whole records if they contained any missing values.
(2) **Imputation of Mean**: Replacing missing values with the mean value of the rest of the values for that feature.
(3) **Imputation of Mode**: Replacing missing values with the modal value of the rest of the values for that feature.

Excision was implemented with the *.dropna* function in the *pandas* data-processing library, whilst both imputation strategies were implemented with the *Imputer* class in *sklearn.preprocessing*. Crucially, the same techniques were not applied to both the training and evaluation data, as that would have altered the size of the holdout set and invalidated the results. Instead, both imputation of mean and imputation of mode were tested on the validation set for each experiment.

In clinical practice, it would be necessary to dose patients even if some data points were absent. It was predicted that imputation techniques would successfully tolerate missing data and perform within 5% of the excision technique.

3.1.3 Experiment 3: New Techniques. The final experiment evaluated new techniques for warfarin dose prediction by comparing them to the best results in the academic literature. The 2015 study by Liu et al. [25] was used to obtain benchmarks. Given the rate of advancement in the field of machine learning – both in technology and methodology – it was predicted that some new techniques would yield superior results to those of Liu et al. when using the same dataset, sample sizes, and parameters. The techniques evaluated included manually optimising 4 promising learning algorithms using *k-fold cross-validation* (see 3.6.2) on the training data, as well as using an *autoML* approach to optimise learning pipelines with genetic programming (see 4).

[5]carbamazepine, phenytoin, or rifampin / rifampicin

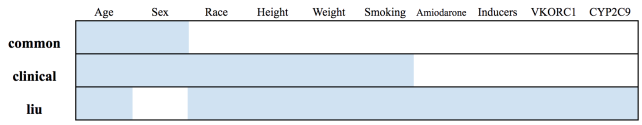

Figure 3: Illustration of the overlapping parameter sets used for experiment 2A.

3.2 Learning Algorithms Used

Two different collections of learning algorithms were used to conduct the above experiments.

3.2.1 Standard Algorithms. This collection of 6 classical machine learning algorithms was used to obtain generalised results in experiments 2A and 2B. All algorithms were implemented through the *scikit-learn* library [31] with hyperparameters set to their default values[6]. The algorithms were linear regression (LR), multi-layer perceptron (MLP), regression tree (RT), support vector regression (SVR), lasso regression (LAS), and Bayesian ridge regression (BR).

3.2.2 Optimised Algorithms. This collection of 11 optimised algorithms was used to establish best predictive performance in experiments 1 and 3. It contained 4 algorithms from *scikit-learn* – lasso Lars (LARS), multi-layer perceptron (MLP), regression tree (RT), and support vector regression (SVR) – which were each manually optimised using 10-fold cross-validation on the training data. This optimisation included both hyperparameter tuning and data preprocessing. The collection also contained 7 machine learning pipelines generated through genetic programming (see 4).

3.3 Datasets

Two datasets of warfarin records were used for this study. The globally-standard IWPC dataset was used primarily for comparing new techniques to those in the literature, whilst the proprietary PathCare dataset was used to evaluate model performance against human experts in a South African context. Analysis revealed that both datasets had similar distributions of weekly warfarin dose, INR, and age. A numerical comparison of the cohorts is provided in the supplementary material (see 10).

3.3.1 IWPC Dataset. The International Warfarin Pharmacogenetics Consortium (IWPC) dataset [48] of 6256 patients has been used in a number of notable studies [23, 25, 38] and is the standard reference point for new methodology in automated warfarin dosing. The dataset was compiled collaboratively and includes data from 22 research groups from 9 countries [23]. As a result, some patient data is missing. When filtering out entries that are missing important clinical values – INR, warfarin dose, weight, height, and age – the resulting dataset contains 4529 records[7].

3.3.2 PathCare Dataset. This dataset was provided by the South African pathology group PathCare specifically for use in this study. As with the IWPC dataset, patients were de-identified, but for legal reasons this dataset was not made publicly available. Unlike the

[6]With the exception of the regression tree, whose default max_depth value of None was set to 10 to constrain the model to the scope of the task.
[7]This concurs with the methodology of Liu et al.'s 2015 study on the same data [25].

IWPC dataset, no pharmacogenetic data was available and very limited clinical data was collected. Most notably, height, weight, and ethnicity data was absent. Also unlike the IWPC case, the PathCare data was provided as a *mySQL* dump file. Extensive data cleaning and reformatting was performed to produce a research-focused *.CSV* file. Although the *mySQL* dump contained over 47000 patient records, only 8985 viable records were found – 69% of which were from patients with only a single record, with the remainder being from the 1135 patients who had multiple visit records. From the total dataset, only 4621 (51%) patient records showed an INR in their target INR range. This dataset also featured 7-day warfarin dosages (in mg) for both taken and prescribed dosages. The weekly total doses were computed for all records. This allowed direct comparison with the IWPC dataset, which lists only the weekly dose.

Record	Colour
1	Green
2	Red
3	NA

vectorisation

Record	Red	Green	Blue
1	0	1	0
2	1	0	0
3	0	0	0

Figure 4: Example of vectorising categorical data into a sparse matrix format. In this instance, the colour category has three classes – red, green, and blue. This illustrates how one parameter (colour) can become three features and missing data is absorbed as a row of zeros. Note that blue is not seen in this dataset, so it must be explicitly encoded into the vectorisation function for this format to generalise.

3.4 Feature Extraction

Machine learning algorithms require a two-dimensional matrix of input values and a one-dimensional vector of target values. To perform the requisite mathematical operations, all values must be numeric. Whilst some data fields – INR values, height in cm, and warfarin dose – were already in continuous format, other data was in categorical or text format and required *vectorisation*. This process generates a feature set larger than the number of parameters provided as input. For instance, the liu et al. parameter set included 9 parameters, but resulted in 58 features.

3.4.1 Vectorisation of Categorical Data. Figure 4 illustrates vectorisation of categorical data using the arbitrary example of colour. In the IWPC dataset, an example of categorical data was the "Race (OMB)" field which has 7 classes – the 7 official OMB values for racial groups in the United States. Vectorising this category into a *sparse matrix* produces 7 features, with each containing a binary digit. This technique is superior to integer indexing as it assumes no relationship between the classes, which results in more accurate models. This is why it is considered best practice [35]. Whilst many preprocessing libraries can vectorise automatically, they are not explicitly aware how many classes exist. Instead, they rely on implicit assumptions from how many unique values they observe. This could result in a different feature-space depending on the dataset used, which meant that trained models could not be evaluated on another dataset unless both contained all possible classes. Moreover, automatic tools ignore missing values but include "NA", "N/A", etc. as categories. To combat this difficulty, the classes for each parameter were explicitly encoded for each dataset. Whilst painstaking, this approach resulted in more accurate representations of the data as well as an appropriate feature-space.

3.4.2 Vectorisation of Text. Text strings – such as lists of which medications patients were taking – posed a similar challenge. Once again, *sparse matrix* format was the preferable output. When processing text strings, the *bag-of-words* technique [35] was used to find the frequency of terms. Important and frequent terms were identified and a new dummy-feature was created for each. Software tools were written to search for these terms in the text strings and insert values into the corresponding feature. In the case of medications, most drugs had at least two different trade names as well as generic names. Moreover, many drug names were difficult to remember or spell, resulting in many similar (but not identical) terms. Finding common mistakes and alternate names for each medication was a slow, manual task, but resulted in a generalised sparse matrix format – an ideal input for a learning algorithm.

3.5 Dataset Splitting

To distinguish a model that overfits the training data from one that accurately fits the underlying relationships, data must be withheld until the end of a study and then used to quantify the general accuracy of that model [35]. This is know as the *holdout method*. For this study, selection was performed at random from the datasets to ensure that the distributions were similar between a training set and its corresponding validation set. The PathCare dataset was scaled[8] and split in the same ratio as the IWPC data, producing datasets of nearly identical size and value distributions[9]. This allowed direct comparison between both datasets with limited confounding factors. The overall splitting protocol is illustrated in figure 5 and detailed below.

The cleaned and transformed PathCare dataset (PATH-whole) contained 8985 data points. For use in experiment 1, this data was split by grouping patients who had visited more than once away from those who only had a single visit. This resulted in a dataset where each *patientID* was unique, named PATH-SV (for single visit), and a smaller dataset of 1135 patients with multiple visits (totalling 2792), named PATH-MV (for multiple visit). For experiment 2, the PATH-whole set was filtered for patient records where the INR was in the patient's target range. This subset of therapeutic patients contained 4621 records and was randomly split 80/20 into the PathCare training (PATH-T) and PathCare validation (PATH-V) sets. This was done *patient*-wise and not *record*-wise (see 3.9.2), ensuring no confounding factors affected model evaluation.

After filtering out records with crucial fields missing[10], 4529 records remained in the IWPC set. These were randomly split 80/20 to create the IWPC training (IWPC-T) and IWPC validation (IWPC-V) sets. This nuanced splitting strategy provided the best possible subsets for training and robust evaluation under each of the experimental conditions.

[8] By excluding the non-therapeutic entries.
[9] This was verified by statistical analysis in the supplementary material (see 10).
[10] Using criteria from Liu et al. [25].

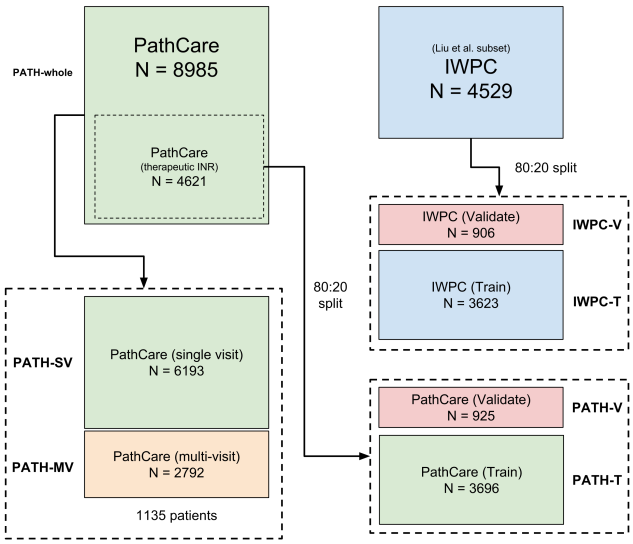

Figure 5: Illustration of dataset splitting protocols for both IWPC and PathCare datasets. To ensure best practice, the holdout sets (IWPC-V and PATH-V) were kept in a withheld directory until final evaluation.

3.6 Training

A wide variety of algorithms (see 3.2) were used to train the predictive models. Most of these were implemented using the *scikit-learn* library [31] in *Python 3.6* and optimised by tuning the hyperparameters supported by the library. In many cases, two instances of the algorithm were used, where the first was manually optimised for the PathCare data and the second for the IWPC data. This allowed the best performances for each algorithm on each dataset to be compared directly.

3.6.1 Feature Selection. Feature selection was performed manually using acquired domain knowledge. In datasets with so few features, all of which were pre-selected for their effect on warfarin metabolism, manual selection was a reasonably efficient and reliable approach. For comparison purposes, the parameter set defined by Liu et al. (see 3.1.2) was used, so no feature selection was necessary. In the case of the PathCare dataset, there were sufficiently few features that manual selection was viable.

3.6.2 Cross-Validation (CV) Methods. Two techniques were used to estimate performance during training. The first was standard *k-fold cross-validation* – where the training set is randomly split into k subsets of identical size and each subset acts as the test set for exactly one iteration [42]. The second was *Monte Carlo cross-validation* – where the dataset is randomly split into train and test subsets for many repetitions [49]. A Monte Carlo CV (MCCV) function was built using the *train_test_split* function in *scikit-learn* and run for 100 iterations with 80/20 splits to find convergent results for the models. These helped to select a value of k for k-fold CV that gave reliable results in a much shorter period of time. k-fold CV was implemented using the *KFold* tool in *scikit-learn*.

For $k = 10$, the performance estimation was fast, accurate, and consistent with MCCV. This combined approach was utilised to facilitate rapid yet robust evaluation of models during manual optimisation.

3.6.3 Preprocessing. Manually-optimised algorithms were enhanced with one of two preprocessing tools available in *scikit-learn*: *StandardScaler* – which removes the mean and scales the data to unit variance – and *RobustScaler* – which centres and scales the data based on percentiles. The automatically-optimised algorithms selected their own preprocessing methods from *sklearn.preprocessing*.

3.7 Clinical and Statistical Metrics

Appropriate metrics were chosen to assess both the clinical and statistical accuracy of the models. INR has a therapeutic range of 2.0-3.0 in most patients [1] and using tighter target ranges for maintenance dosing does not achieve any therapeutic advantage [28]. The chosen metrics accounted for that, and were consistent with metrics used in related studies – allowing direct comparisons.

(1) **Mean absolute error (MAE)** is widely used across warfarin prediction studies [12, 15, 25, 26, 38, 39, 43, 50].

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{1}$$

where y_i is the actual dose and x_i is the predicted dose.

(2) **Percentage of patients with dose estimates within 20% of the actual dose (PW20)** was used by a number of notable studies [12, 23, 25, 50]. It reflects the fact that being within 0.5 points of the target INR is clinically sufficient.

$$PW20 = \frac{\sum_{i=1}^{n} f(p_i)}{n} \tag{2}$$

where $f(p_i)$ for patient p_i is 1 if $0.8y_i < x_i < 1.2y_i$, else 0; x_i is the predicted dose, and y_i is the true dose.

(3) **R^2-value (coefficient of determination)** is the gold standard for statistical comparison and was utilised in numerous studies [5, 12, 16, 25]. It highlights how well the model fits the data as a whole. R^2 was used to verify that the MAE and PW20 results were legitimate, and as a third metric to differentiate algorithms with similar performance.

3.8 Evaluation Protocol

For experiment 1, human performance was estimated using PATH-MV (see 3.9.1). This was compared with the 11 optimised models trained on PATH-T and evaluated on the withheld PATH-V set. For experiments 2A and 2B, the 6 standard models were trained on PATH-T and IWPC-T and then evaluated on the withheld PATH-V and IWPC-V sets respectively. This was repeated for each parameter set and each missing data treatment, as detailed in section 3.1.2. For experiment 3, the 11 optimised models were trained and evaluated using 80/20 *Monte Carlo CV* with 100 iterations on the filtered IWPC dataset. This was similar to the 100 rounds of 80/20 re-sampling performed by Liu et al. [25] and used the same filtering parameters, allowing for a direct performance comparison across studies. To determine if the results generalised, the *holdout* method was used – models were trained on IWPC-T and evaluated on IWPC-V. All

MAE and PW20 scores from each experiment were verified by examining the R^2 values[11].

3.9 Methodological Challenges and Solutions

3.9.1 Estimation of Human Performance. Experiment 1 compared the performance of clinicians to models on a subset of the PathCare dataset that contained patients with multiple visits (PATH-MV). Because no clinical experiment could be run whereby models dosed real patients, predictions were compared to final therapeutic doses. Because those final therapeutic doses were only available in cases where the patient was successfully dosed, estimated therapeutic doses were needed for the remaining cases. A novel technique was developed to impute those estimates with random sampling from a *Gaussian* distribution with the same mean and variance as the PathCare data[12]. As a result of the *central limit theorem*, a good estimate emerged when averaged. This method ensured a fair evaluation by eliminating the selection bias on the data used to measure human performance.

3.9.2 Patient-wise Splitting. The PATH-V and PATH-T datasets were originally produced by splitting *record*-wise in the same way as the IWPC dataset. Because the PathCare data contained multiple visits for some patients, the PATH-V and PATH-T sets were processed to ensure that no patients had records in both sets[13]. A novel technique was used to handle records of patients that appeared in both sets by swapping them out with a randomly-chosen record from the other set. This was repeated until no *patientIDs* were seen across both sets. The swapping procedure emulated the effect of splitting the PathCare data *patient*-wise, but without the arduous task of ensuring that the 80/20 ratio was preserved.

3.10 Software Architecture

The project was written in *Python 3.6* with *PEP8* conventions, full documentation, and a style focused on modularity. Functions and classes were removed if they became redundant and all code was regularly refactored. Code readability was a priority, as it lets experts in the domain of warfarin dosing more easily understand (and augment) the program logic – allowing those with only an intermediate understanding of *Python* to conduct further research. The code was open-sourced at `github.com/gianlucatruda/warfit` and welcomes improvements.

4 OPTIMISATION WITH GENETIC PROGRAMMING

A Tree-based Pipeline Optimisation Tool (TPOT) was used to generate high-performing pipelines through genetic programming [30]. Cleaned versions of PATH-T and IWPC-T were given as input and many generations of supervised learning yielded the best performers – optimised meta-algorithms that would likely never have been found through manual implementation and tuning. TPOT accepts bespoke scoring functions as its fitness function. The functions for

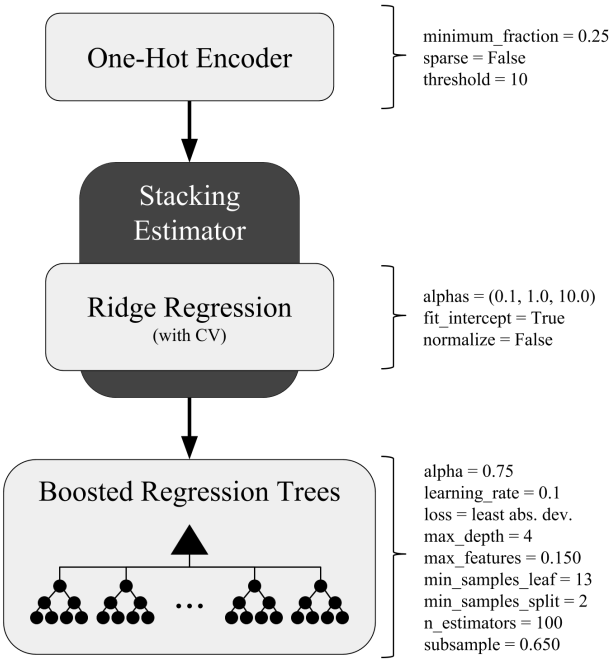

Figure 6: Graphical depiction of the best performing learning algorithm, pipeline 5 (PL5), the result of optimisation through genetic programming with TPOT.

PW20 and MAE were tested, as was a hybrid of the two:

$$hybrid = \frac{PW20}{MAE^2} \tag{3}$$

Many instances of TPOT were run using different evolutionary parameters. The number of generations ranged from 50 to 10000, the number of offspring from 5 to 100, and the k used in k-fold CV from 5 to 30. Over all these instances, the 7 best performers were exported and evaluated against human-developed algorithms.

4.1 Pipeline 5: The Top Performer

Pipeline 5 (PL5) first performs one-hot encoding on the input data, then passes that result to a stacking estimator – which adds synthetic features derived from ridge regression with cross-validation – and then provides the resulting data as the input to an ensemble of boosted regression trees using the *least absolute deviation* loss function. This is illustrated in figure 6, which also displays the hyperparameters.

5 RESULTS AND DISCUSSION

This section details how learning algorithms were able to perform comparably with human experts on South African data. It then demonstrates the efficacy of imputation for handling missing data and the positive relationship between feature-space and model accuracy. Finally, it is shown that a TPOT pipeline outperformed the best models in the published literature. Full results are provided in the supplementary material (see 10), including R^2 verification.

[11] Full details for all metrics are shown in the supplementary material (see 10).
[12] Details of this method can be found in the online supplementary material (see 10).
[13] as that may have conferred unrealistic meta-knowledge to the learning algorithms

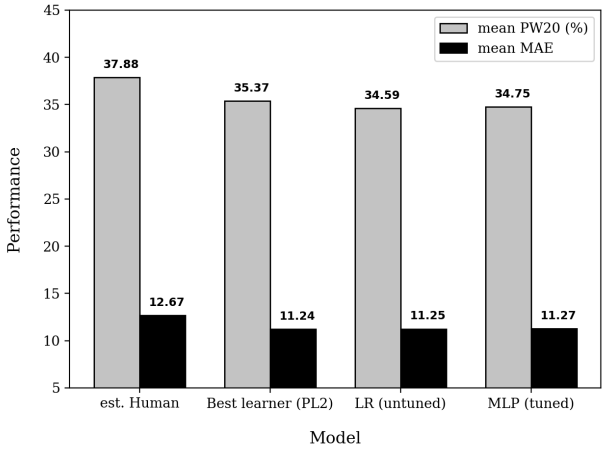

Figure 7: Comparison of estimated human dosing performance compared to the best learning pipeline (PL2), an untuned linear regression algorithm (LR), and a tuned multilayer perceptron (MLP). The human performance was estimated using the PATH-MV dataset and averaged over 50 iterations.

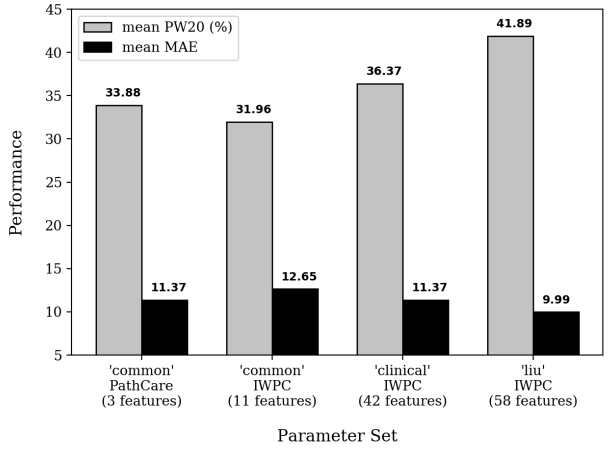

Figure 8: Comparison of mean performance across parameter sets for a collection of 6 machine learning algorithms – LR, RT, SVR, MLP, LASSO, BR.

5.1 Models and Human Experts

Figure 7 compares estimated human performance with the mean performance of three of the 17 models trained on PATH-T and validated on PATH-V[14]. The best-performing parameter set[15] and *mean imputation* for missing values were used for both training and validation. The figure shows only the best overall performer (pipeline 2), the simplest algorithm (untuned linear regression), and an example of a multi-layer perceptron with manually-tuned hyperparameters. The human estimate outperformed the best learner by only 7.1% in terms of PW20, but the learner had an MAE 11.3% better. This supports the initial hypothesis that learned models could measure up to human experts. These findings suggest that even with the limited format of the current PathCare database, learning algorithms can perform as well as human experts when it comes to prescribing therapeutic maintenance doses. Whilst more customised algorithms (like PL2) performed slightly better, even an off-the-shelf linear regression algorithm performed favourably compared with human estimates.

5.2 The Effects of Data-Manipulations

Figure 8 compares the mean performance of the 6 standard learning algorithms across the *common*, *clinical*, and *liu* parameter sets outlined in section 3.1.2. *Mean imputation* was used to handle all missing data in both training and validation sets and results were averaged over 20 iterations. Despite the *common* parameter set containing only age in years and sex, it resulted in different feature sizes in the PathCare and IWPC datasets. This was because the IWPC used 9 age categories instead of a continuous value like in the PathCare set. Despite this, the *common* parameter set resulted in lower mean performance on the IWPC dataset – 5.7% worse in terms of PW20 and 11.3% worse in terms of MAE. This is most likely due to the fact that the PATH-T and PATH-V sets were taken from only the therapeutic cases in the PathCare data, which may have made data points slightly more consistent than in the IWPC set. A minor discrepancy across datasets was to be expected and was quite small relative to the other results. Overall, the learning algorithms performed similarly across both the PathCare and IWPC datasets, which stands to validate the methodology. The *clinical* and *liu* parameter sets resulted in far more features and increased performance in terms of both PW20 and MAE. The *liu* parameter set was the only one to include pharmacogenetic parameters. This caused a dramatic increase in performance over the *common* parameter set – 29.5% in terms of PW20 and 21.0% in terms of MAE. This supports the value of pharmacogenetic factors in predicting maintenance doses. In general, as relevant parameters were added, the predictive performance increased, suggesting a positive relationship exists. This supports the hypothesis in section 3.1.2. Moreover, these results suggest that even without pharmacogenetic data, easily-obtained clinical data such as height, weight, race and smoking history can boost performance by at least 10%.

Figure 9 compares treatments for missing data for the 6 standard learning algorithms. All models were trained on IWPC-T using the *liu* parameter set (58 features) and evaluated on IWPC-V[16]. Both imputation treatments slightly outperformed the excision treatment in terms of both PW20 and MAE. Imputed *mean* performed slightly better than imputed *mode* and was used throughout the other experiments[17]. These results support the hypothesis that imputation

[14] With results averaged over 10 iterations.

[15] path2 = sex; year of birth; and sparse vectors for aspirin use, paracetamol use, amiodarone use, atrial fibrillation, deep vein thrombosis, and valve replacement.

[16] To keep the size and contents of the validation set consistent, the same treatment was used on missing data in the validation set each time, regardless of the treatment method being investigated.

[17] It should be noted that imputed mode outperforms imputed mean when imputed mode is also used on the validation set, but the difference is less significant. It would seem that either treatment can be used, but that best performance comes when using the same treatment on both the validation set and the training set.

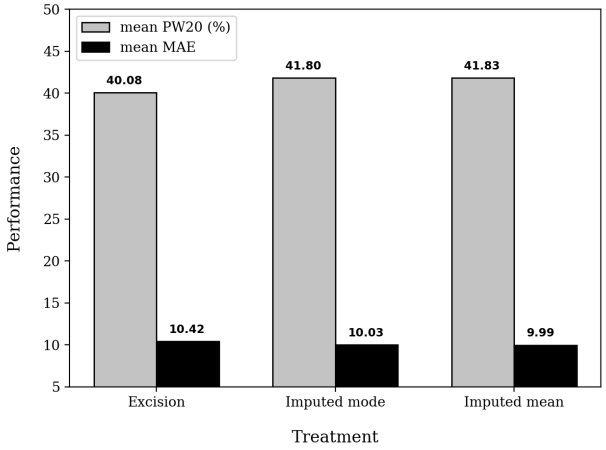

Figure 9: Comparison of mean performance of treatments for missing data for the 6 standard learning algorithms – LR, RT, SVR, MLP, LASSO, BR.

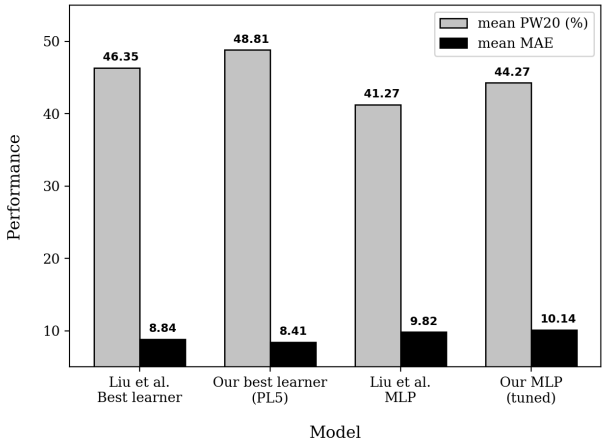

Figure 10: Comparison of Liu et al. best performer (MARS) and multi-layer perceptron with this study's best performer and multi-layer perceptron. PL5 set a new performance benchmark for warfarin dosing. Liu et al. results were taken from table 2 of their 2015 paper [25].

could be used to handle missing data without compromising accuracy. This bodes well for clinical implementation, as imputation allows for more robust prediction models.

The difference between excision and imputations treatments was likely this small because the datasets and feature sizes were relatively small. This difference could be expected to scale as either the features or examples become more numerous. A likely reason for the similar performance between imputation methods is that the minor variations in generated values became almost unnoticeable when models preprocessed the data during training and evaluation.

5.3 Performance of New Techniques

Figure 10 compares the best and typical learning algorithms in the Liu et al. study [25] with the best and typical algorithms from this study. The same data and parameter set was used in each case, with 100 rounds of *Monte Carlo CV*. This study's top performing learning algorithm was pipeline 5 (PL5), generated through TPOT (see 4). It outperformed Liu et al.'s multiple adaptive regression splines (MARS) learner by 5.3% in terms of PW20 and 4.9% in terms of MAE. This supports the hypothesis that contemporary techniques could improve the best results on IWPC data. The *holdout method* verified that these results generalised to within 5% on the withheld IWPC-V data[18].

This study's tuned multi-layer perceptron (MLP) outperformed Liu et al.'s MLP by 7.3% in terms of PW20, but performed 3.2% less well in terms of MAE. This suggests that manually tuning hyperparameters will tend to favour one performance metric (such as PW20) at the expense of others. With genetically-generated pipelines and hybrid metrics, however, new learning algorithms can be created without expert domain knowledge. Moreover, the *autoML* models can significantly outperform those of manually-tuned algorithms. This, and the superior performance of PL5, supports the efficacy of *autoML* techniques in warfarin dosing.

[18]See the full results and verification in the supplementary material (see 10).

5.4 Replicability

Replication is of the utmost importance in the scientific process. Accordingly, this project facilitated easy replication by open-sourcing the code. Given that the IWPC dataset is in the public domain[19] and the code for this project is fully-documented, replicating most of these results should be trivial. This paper also details all methodology employed. Unfortunately, the PathCare dataset remains proprietary, but given the potential impact of these findings, PathCare is encouraged to replicate these results using the provided resources as guidelines. The author can be contacted via the project repository for further details where necessary.

6 ETHICAL AND PROFESSIONAL ISSUES

PathCare de-identified the individuals in their dataset, rendering it impossible for anyone to link data points back to the patients from which they came. The dataset remains private. Ethics clearance for this study was obtained via the *Faculty of Science Research Ethics Committee*. In the event of publication, the guidelines of the UCT *Authorship Practices Policy* will be observed, with contributing parties being acknowledged for their share of work, and intellectual property derived from the project being subject to UCT's *Intellectual Property Policy*. PathCare provided the dataset with no conditions beyond ethics approval. The author declares no conflict of interest.

7 LIMITATIONS

Most fields in the PathCare database were of type *varChar*, which allows any textual input. This made erroneous formatting and misplaced data ubiquitous, which reduced the number of viable records and may have reduced the overall performance. Moreover, the limited number of visits for most patients in the dataset suggest that many individuals mistakenly had multiple *patientIDs* assigned

[19]www.pharmgkb.org/downloads

to them, with an impact on the results of both the models and human estimates. The database currently does not accommodate important clinical information such as height, weight, and race – all of which would have drastically improved dosing accuracy.

The IWPC dataset is compiled from 22 research groups from nine countries, each with different protocols and equipment. This results in noisy data and missing values, lowering the predictive accuracy of models. In clinical implementation, increased accuracy is likely (given sufficient data). It is also known that the impact of CYP2C9 and VKORC1 genes varies across races [25]. Current data is mostly derived from "White" and "Asian" racial groups in developed nations, so it is likely that current pharmacogenetic implementations impart bias upon models.

8 CONCLUSIONS AND FUTURE WORK

The experiments demonstrated that learning algorithms can produce models at least as effective as human experts at prescribing warfarin maintenance doses. This was observed in the current PathCare dataset – which is relatively small and poorly structured – and even using the simplest techniques, such as linear regression. If provided with richer, more comprehensive data, learning algorithms would likely outperform haemetopathologists and pave a path to automated dosing.

Whilst simpler techniques are outperformed by more advanced ones, the difference was marginal in datasets with smaller feature-space. As the number of relevant parameters was increased, the performance advantage of more complex learning algorithms also increased. This suggests that as more clean data becomes available, advanced algorithms will become necessary to fully utilise the available data.

Unfortunately, data cleaning and formatting consume disproportionate amounts of time and require sacrifices to the quality of the training data. One of the best routes to increasing model performance would be to implement stricter data collection and review protocols. This would increase the ratio of usable data and likely result in more effective dosing.

Currently, PathCare collects very few of the clinical metrics relevant to accurate warfarin dosing. Implementing new policies and database schema to collect height, weight, race, and smoking status would be of relative ease, but has been shown in both this and many other studies to drastically improve the dosing accuracy of models. Many studies have observed the benefits of pharmacogenetic data in warfarin dosing, and this study supports those findings. Whilst this is valuable information, the requisite tests are still highly inaccessible to South Africans. The resources required for pharmacogenetic warfarin dosing are not yet justified by the performance increases in the context of South African healthcare.

This study observed that imputation methods were an effective means of dealing with missing data. This is especially important in warfarin dose prediction, where datasets are both small and incomplete. By imputing missing values instead of removing the entire record, more data is available for training, which boosts performance. Moreover, the ability to impute missing features is essential to clinical application of these models, as it allows them to dose future patients even if some parameters are not available – which is a frequent occurrence in clinical practice.

This study found that automatic machine learning techniques – in this case optimisation through genetic programming – were an effective method of producing accurate models with limited domain knowledge. This eliminates the need for machine learning expertise, which drastically improves the resource efficiency and availability of dose prediction. The automatically-generated learning pipelines produced were not only simpler to attain than manually-tuned algorithms, but also performed better. If this trend is not unique to warfarin dosing, it suggests that *autoML* is a promising method for the future of automated dosing in general, which is of huge importance to computer science.

In their totality, the results of this study highlight how machine learning techniques offer immense value to warfarin dosing in South Africa. In the short term, organisations like PathCare could begin implementing models as an aid to the human experts. This is likely to reduce error rates and result in more effective dosing, whilst creating a *human-in-the-loop* architecture that can be used to enhance model performance going forward. In the medium term, more feature-rich data would be collected and model performance would improve accordingly. This would help transition toward a future of automated warfarin maintenance dosing, which can free up haematopathologists to address other pressing patient issues. During that transition, improved predictions would reduce the time required to get patients into the therapeutic range and reduce the incidence of misdosing.

In future research, the findings of this study should be replicated in other automated dosing problems. Some aspects of the findings may be specific to warfarin dosing or to the datasets used to train and evaluate the models. It is therefore crucial to establish how general the findings are to the domain of drug dosing. There is also much work to be done in creating a framework for testing new techniques on the IWPC dataset, including robust tools and evaluation metrics. This could be achieved by extending this project's open-sourced code, eliminating the need for future researchers to re-create existing tools, and keeping conditions consistent – allowing rapid evaluation of new techniques on a level playing field. Ultimately, the greatest advances can still be made in compiling more comprehensive warfarin dosing datasets from more diverse cohorts, which can be used to re-evaluate the learning algorithms and train even better models.

9 ACKNOWLEDGEMENTS

The author wishes to thank Associate Professor Patrick Marais for supervising this project and providing insight and guidance throughout. Thanks are also extended to PathCare – specifically to Gill Land, Dr Illse Louw, and Marike Ubbink – for providing this dataset and practical insights.

10 SUPPLEMENTARY MATERIAL

A document with supplementary material is provided online at `http://bit.ly/warfit_supp`. It contains raw data, more graphics, and details of some techniques.

REFERENCES

[1] G. W. Albers, J. E. Dalen, A. Laupacis, W. J. Manning, P. Petersen, and D. E. Singer. 2001. Antithrombotic therapy in atrial fibrillation. *Chest* 119, 1 SUPPL. (2001). https://doi.org/10.1378/chest.119.1

[2] Jeffrey L. Anderson, Benjamin D. Horne, Scott M. Stevens, Amanda S. Grove, Stephanie Barton, Zachery P. Nicholas, Samera F.S. Kahn, Heidi T. May, Kent M. Samuelson, Joseph B. Muhlestein, and John F. Carlquist. 2007. Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. *Circulation* 116, 22 (2007), 2563–2570. https://doi.org/10.1161/CIRCULATIONAHA.107.737312

[3] Wolfgang Banzhaf, Peter Nordin, Robert E Keller, and Frank D Francone. 1998. *Genetic programming: an introduction*. Vol. 1. Morgan Kaufmann San Francisco.

[4] Erdal Cosgun, Nita A. Limdi, and Christine W. Duarte. 2011. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. *Bioinformatics* 27, 10 (2011), 1384–1389. https://doi.org/10.1093/bioinformatics/btr159

[5] Erdal Cosgun, Nita A. Limdi, and Christine W. Duarte. 2011. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. *Bioinformatics* 27, 10 (2011), 1384–1389. https://doi.org/10.1093/bioinformatics/btr159

[6] Khagendra Dahal Dahal. 2014. Meta-analysis of Randomized Controlled Trials of Genotype-guided versus Standard Dosing of Warfarin. *CHEST Journal* (2014), 1–27. https://doi.org/10.1007/s00246-002-9361-x

[7] Pedro Domingos. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books. https://books.google.co.za/books?id=glUtrgEACAAJ

[8] Stephanie Forrest, ThanhVu Nguyen, Westley Weimer, and Claire Le Goues. 2009. A genetic programming approach to automated software repair. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*. ACM, 947–954.

[9] FÃĪlix-Antoine Fortin, FranÃğois-Michel De Rainville, Marc-AndrÃĪ Gardner, Marc Parizeau, and Christian Gagné. 2012. DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research* 13, Jul (2012), 2171–2175.

[10] Erik M Fredericks and Betty H C Cheng. 2013. Exploring automated software composition with genetic programming. In *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*. ACM, 1733–1734.

[11] B. F. Gage, C. Eby, J. A. Johnson, E. Deych, M. J. Rieder, P. M. Ridker, P. E. Milligan, G. Grice, P. Lenzini, A. E. Rettie, C. L. Aquilante, L. Grosso, S. Marsh, T. Langaee, L. E. Farnett, D. Voora, D. L. Veenstra, R. J. Glynn, A. Barrett, and H. L. McLeod. 2008. Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. *Clinical Pharmacology and Therapeutics* 84, 3 (2008), 326–331. https://doi.org/10.1038/clpt.2008.10

[12] Enzo Grossi, Gian Marco Podda, Mariateresa Pugliano, Silvia Gabba, Annalisa Verri, Giovanni Carpani, Massimo Buscema, Giovanni Casazza, and Marco Cattaneo. 2014. Prediction of optimal warfarin maintenance dose using advanced artificial neural networks. *Pharmacogenomics* 15, 1 (2014), 29–37. https://doi.org/10.2217/pgs.13.212

[13] Takumi Harada, Noritaka Ariyoshi, Hitoshi Shimura, Yasunori Sato, Iichiro Yokoyama, Kaori Takahashi, Shin Ichi Yamagata, Mizuho Imamaki, Yoshio Kobayashi, Itsuko Ishii, Masaru Miyazaki, and Mitsukazu Kitada. 2010. Application of Akaike information criterion to evaluate warfarin dosing algorithm. *Thrombosis Research* 126, 3 (2010), 183–190. https://doi.org/10.1016/j.thromres.2010.05.016

[14] Gregory S Hornby, Jason D Lohn, and Derek S Linden. 2011. Computer-automated evolution of an X-band antenna for NASA's space technology 5 mission. *Evolutionary computation* 19, 1 (2011), 1–23.

[15] Ya Han Hu, Fan Wu, Chia Lun Lo, and Chun Tien Tai. 2012. Predicting warfarin dosage from clinical data: A supervised learning approach. *Artificial Intelligence in Medicine* 56, 1 (2012), 27–34. https://doi.org/10.1016/j.artmed.2012.04.001

[16] Hussain A. Isma'Eel, George E. Sakr, Robert H. Habib, Mohamad Musbah Almedawar, Nathalie K. Zgheib, and Imad H. Elhajj. 2014. Improved accuracy of anticoagulant dose prediction using a pharmacogenetic and artificial neural network-based method. *European Journal of Clinical Pharmacology* 70, 3 (3 2014), 265–273. https://doi.org/10.1007/s00228-013-1617-2

[17] J. A. Johnson, K. E. Caudle, L. Gong, M. Whirl-Carrillo, C. M. Stein, S. A. Scott, M. T. Lee, B. F. Gage, S. E. Kimmel, M. A. Perera, J. L. Anderson, M. Pirmohamed, T. E. Klein, N. A. Limdi, L. H. Cavallari, and M. Wadelius. 2017. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin Dosing: 2017 Update. *Clinical Pharmacology and Therapeutics* 102, 3 (2017), 397–404. https://doi.org/10.1002/cpt.668

[18] Daniel E. Jonas and Howard L. McLeod. 2009. Genetic and clinical factors relating to warfarin dosing. *Trends in Pharmacological Sciences* 30, 7 (2009), 375–386. https://doi.org/10.1016/j.tips.2009.05.001

[19] Sue Jowett, S. Bryan, L. Poller, A. M.H.P. Van Den Besselaar, F. J.M. Van Der Meer, G. Palareti, C. Shiach, A. Tripodi, M. Keown, S. Ibrahim, G. Lowe, M. Moia, A. G. Turpie, and J. Jespersen. 2009. The cost-effectiveness of computer-assisted anticoagulant dosage: Results from the European Action on Anticoagulation (EAA) multicentre study. *Journal of Thrombosis and Haemostasis* 7, 9 (9 2009), 1482–1490. https://doi.org/10.1111/j.1538-7836.2009.03508.x

[20] David M. Keeling, Trevor Baglin, Campbell Tait, Henry Watson, David Perry, Caroline Baglin, Steve Kitchen, and Michael Makris. 2011. Guidelines on oral anticoagulation with warfarin - fourth edition. *British Journal of Haematology* 154, 3 (2011), 311–324. https://doi.org/10.1111/j.1365-2141.2011.08753.x

[21] Y. K. Kim, R. Nieuwlaat, S. J. Connolly, S. Schulman, K. Meijer, N. Raju, S. Kaatz, and J. W. Eikelboom. 2010. Effect of a simple two-step warfarin dosing algorithm on anticoagulant control as measured by time in therapeutic range: A pilot study. *Journal of Thrombosis and Haemostasis* 8, 1 (2010), 101–106. https://doi.org/10.1111/j.1538-7836.2009.03652.x

[22] T B L Kirkwood and S M Lewis. 1983. Requirements for thromboplastins and plasma used to control oral anticoagulant therapy. *WHO Tech Rep Ser* 687 (1983), 81–99.

[23] Altman R.B. Eriksson B.F. Klein, T.E. 2009. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. *Archives of Internal Medicine* 360, 8 (2009), 753–764. https://doi.org/10.1056/NEJMoa0809329.Estimation

[24] G. Le Gal, M. Carrier, S. Tierney, H. Majeed, M. Rodger, and P. S. Wells. 2010. Prediction of the warfarin maintenance dose after completion of the 10 mg initiation nomogram: Do we really need genotyping? *Journal of Thrombosis and Haemostasis* 8, 1 (2010), 90–94. https://doi.org/10.1111/j.1538-7836.2009.03676.x

[25] Rong Liu, Xi Li, Wei Zhang, and Hong Hao Zhou. 2015. Comparison of nine statistical model based warfarin pharmacogenetic dosing algorithms using the racially diverse international warfarin pharmacogenetic consortium cohort database. *PLoS ONE* 10, 8 (2015). https://doi.org/10.1371/journal.pone.0135784

[26] Yu Liu, Jie Yang, Qiang Xu, Bin Xu, Lei Gao, Yuxiao Zhang, Yan Zhang, Hongjuan Wang, Caiyi Lu, Yusheng Zhao, and Tong Yin. 2012. Comparative performance of warfarin pharmacogenetic algorithms in Chinese patients. *Thrombosis Research* 130, 3 (2012), 435–440. https://doi.org/10.1016/j.thromres.2012.02.003

[27] S. McDonald, C. Xydeas, and P. Angelov. 2008. A retrospective comparative study of three data modelling techniques in anticoagulation therapy. *BioMedical Engineering and Informatics: New Development and the Future - Proceedings of the 1st International Conference on BioMedical Engineering and Informatics, BMEI 2008* 1 (2008), 219–225. https://doi.org/10.1109/BMEI.2008.298

[28] D. J. Meier, S. Seva, and William P. Fay. 2007. A comparison of anticoagulation results of patients managed with narrow vs. standard international normalized ratio target ranges [6]. *Journal of Thrombosis and Haemostasis* 5, 6 (2007), 1332–1334. https://doi.org/10.1111/j.1538-7836.2007.02561.x

[29] Liyan Miao, Jian Yang, Chenrong Huang, and Zhenya Shen. 2007. Contribution of age, body weight, and CYP2C9 and VKORC1 genotype to the anticoagulant response to warfarin: Proposal for a new dosing regimen in Chinese patients. *European Journal of Clinical Pharmacology* 63, 12 (11 2007), 1135–1141. https://doi.org/10.1007/s00228-007-0381-6

[30] Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, and Jason H. Moore. 2016. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. (2016), 485–492. https://doi.org/10.1145/2908812.2908918

[31] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[32] Munir Pirmohamed, Girvan Burnside, Niclas Eriksson, Andrea L. Jorgensen, Cheng Hock Toh, Toby Nicholson, Patrick Kesteven, Christina Christersson, Bengt Wahlström, Christina Stafberg, J. Eunice Zhang, Julian B. Leathart, Hugo Kohnke, Anke H. Maitland-van der Zee, Paula R. Williamson, Ann K. Daly, Peter Avery, Farhad Kamali, and Mia Wadelius. 2013. A Randomized Trial of Genotype-Guided Dosing of Warfarin. *New England Journal of Medicine* 369, 24 (2013), 2294–2303. https://doi.org/10.1056/NEJMoa1311386

[33] L. Poller. 2004. International Normalized Ratios (INR): The first 20 years. *Journal of Thrombosis and Haemostasis* 2, 6 (2004), 849–860. https://doi.org/10.1111/j.1538-7836.2004.00775.x

[34] Leon Poller, M. Keown, S. Ibrahim, G. Lowe, M. Moia, A. G. Turpie, C. Roberts, A. M.H.P. Van Den Besselaar, F. J.M. Van Der Meer, A. Tripodi, G. Palareti, C. Shiach, S. Bryan, M. Samama, M. Burgess-Wilson, A. Heagerty, P. Maccallum, D. Wright, and J. Jespersen. 2008. An international multicenter randomized study of computer-assisted oral anticoagulant dosage vs. medical staff dosage. *Journal of Thrombosis and Haemostasis* 6, 6 (6 2008), 935–943. https://doi.org/10.1111/j.1538-7836.2008.02959.x

[35] Sebastian Raschka. 2015. Python machine learning. Packt Publishing Ltd, Chapter 4, 6, 7, 8.

[36] H. Schelleman, J. Chen, Z. Chen, J. Christie, C. W. Newcomb, C. M. Brensinger, M. Price, A. S. Whitehead, C. Kealey, C. F. Thorn, F. F. Samaha, and S. E. Kimmel. 2008. Dosing algorithms to predict warfarin maintenance dose in Caucasians and African Americans. *Clinical Pharmacology and Therapeutics* 84, 3 (2008), 332–339. https://doi.org/10.1038/clpt.2008.101

[37] Elizabeth A. Sconce, Tayyaba I. Khan, Hilary A. Wynne, Peter Avery, Louise Monkhouse, Barry P. King, Peter Wood, Patrick Kesteven, Ann K. Daly, and Farhad Kamali. 2005. The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: Proposal for a new dosing regimen. *Blood* 106, 7 (2005), 2329–2333. https://doi.org/10.1182/blood-2005-03-1108

[38] Ashkan Sharabiani, Adam Bress, Elnaz Douzali, and Houshang Darabi. 2015. Revisiting warfarin dosing using machine learning techniques. *Computational and Mathematical Methods in Medicine* 2015 (2015). https://doi.org/10.1155/2015/

560108

[39] Ashkan Sharabiani, Houshang Darabi, Adam Bress, Larisa Cavallari, Edith Nutescu, and Katarzyna Drozda. 2013. Machine learning based prediction of warfarin optimal dosing for African American patients. In *Automation Science and Engineering (CASE), 2013 IEEE International Conference on*. 623–628. https://doi.org/10.1109/CoASE.2013.6653999

[40] Idit Solomon, Nitsan Maharshak, Gal Chechik, Leonard Leibovici, Aharon Lubetsky, Hillel Halkin, David Ezra, and Nachman Ash. 2004. Applying an artificial neural network to warfarin maintenance dose prediction. *Israel Medical Association Journal* 6, 12 (2004), 732–735.

[41] Lee Spector, David M Clark, Ian Lindsay, Bradford Barr, and Jon Klein. 2008. Genetic programming for finite algebras. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*. ACM, 1291–1298.

[42] M. Stone. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. , 111–147 pages. https://doi.org/10.2307/2984809

[43] S. L. Tan, Z. Li, G. B. Song, L. M. Liu, W. Zhang, J. Peng, T. Zhang, F. F. Jia, G. Zhou, H. H. Zhou, and X. M. Zhou. 2012. Development and comparison of a new personalized warfarin stable dose prediction algorithm in Chinese patients undergoing heart valve replacement. *Pharmazie* 67, 11 (2012), 930–937. https://doi.org/10.1691/ph.2012.2633

[44] Gary Tse, Mengqi Gong, Guangping Li, Sunny Hei Wong, William K. K. Wu, Wing Tak Wong, Leonardo Roever, Alex Pui Wai Lee, Gregory Y. H. Lip, Martin C. S. Wong, and Tong Liu. 2018. Genotype-guided warfarin dosing vs . conventional dosing strategies: a systematic review and meta-analysis of randomized controlled trials. *British Journal of Clinical Pharmacology* (2018). https://doi.org/10.1111/bcp.13621

[45] Mia Wadelius, Leslie Y. Chen, Niclas Eriksson, Suzannah Bumpstead, Jilur Ghori, Claes Wadelius, David Bentley, Ralph McGinnis, and Panos Deloukas. 2007. Association of warfarin dose with genes involved in its action and metabolism. *Human Genetics* 121, 1 (2007), 23–34. https://doi.org/10.1007/s00439-006-0260-8

[46] Mia Wadelius, Leslie Y. Chen, Jonatan D. Lindh, Niclas Eriksson, Mohammed J.R. Ghori, Suzannah Bumpstead, Lennart Holm, Ralph McGinnis, Anders Rane, and Panos Deloukas. 2009. The largest prospective warfarin-treated cohort supports genetic forecasting. *Blood* 113, 4 (2009), 784–792. https://doi.org/10.1182/blood-2008-04-149070

[47] P S Wells, A M Holbrook, N R Crowther, and J Hirsh. 1994. Interactions of warfarin with drugs and food. , 676–683 pages. https://doi.org/10.7326/0003-4819-121-9-199411010-00009

[48] M Whirl-Carrillo, EM McDonogh, J Herbet, L Gong, K Sangkuhl, C Thotn, R Altman, and E Klein. 2012. Pharmacogenomics Knowledge for Personlized Medicine. *Clinical Pharmacology and Therpeutics* 92, 4 (2012), 414–417. https://doi.org/10.1038/clpt.2012.96.Pharmacogenomics

[49] Qing-Song Xu and Yi-Zeng Liang. 2001. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* 56, 1 (4 2001), 1–11. https://doi.org/10.1016/S0169-7439(00)00122-2

[50] Qin Zhou, Joey Kwong, Jie Chen, Wenzhe Qin, Jin Chen, and Li Dong. 2014. Use of artificial neural network to predict warfarin individualized dosage regime in Chinese patients receiving low-intensity anticoagulation after heart valve replacement. *International Journal of Cardiology* 176, 3 (10 2014), 1462–1464. https://doi.org/10.1016/j.ijcard.2014.08.062