Improving Automated Warfarin Dosing for South Africans using Machine Learning

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

Neville Varney-Horwitz
VRNNEV001
University of Cape Town
vrnnev001@myuct.ac.za

CCS CONCEPTS

• **Applied computing** → **Consumer health**; *Health care information systems*; Genetics; • **Computing methodologies** → *Supervised learning*; *Machine learning approaches*;

1 PROJECT DESCRIPTION

Many individuals suffer from abnormalities in blood coagulation which can lead to obstructive blood clots, strokes, and heart attacks. The standard method for treating these conditions is the use of anticoagulant drugs, such as warfarin. Whilst oral warfarin treatment is extremely 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 [13, 35]. This makes the task of accurately dosing warfarin a highly individualised endeavour.

To simplify and standardise the process of anticoagulant monitoring, the World Health Organisation established the international normalised ratio (INR) as a universal reference value [18], with a recommended therapeutic range between 2.0 and 3.0 for most patients [1]. Dosing protocols have attempted to formalise the dosing procedure [12, 15], and software tools exist to assist clinicians [14, 16], but the high individual variability of warfarin, and the risk of severe bleeding, make the development of more accurate dosing methods an ongoing priority.

Many studies [7, 8, 10, 11, 19, 20, 22, 28–30, 37] have looked at applying machine learning techniques to the problem of individualised warfarin dosing, but the datasets are often small and restricted to specific population groups. To address this, we propose a series of experiments to determine the value of training warfarin dosing models on a local dataset. This is possible with access to comprehensive warfarin records provided by the pathology group PathCare. We suggest it may be possible to use this dataset to construct a model for warfarin dosing that offers real benefits to patients in South Africa.

Artificial neural networks (ANNs) are one of the most widespread and successful [20] approaches to warfarin dose prediction. In some studies [10], ensemble techniques like bagging and voting are used to improve the robustness and accuracy of ANNs for warfarin dosing. We term this class of algorithms *wide*, as they utilise many simple ANNs to collectively produce a consensus prediction. This would be in contrast to a *deep* algorithm, which would make use of only a single, highly-layered ANN for warfarin dose prediction.

UCT, CSC4000W, May 2018
2018.

Despite the success of ANNs in warfarin dosing research, there are no notable attempts to apply deep learning – using ANNs with many hidden layers – to the problem. We propose a series of experiments to compare the performance of deep and wide algorithms both to one another, and to conventional techniques like linear regression (LR).

The combined evaluation of models trained on local datasets using deep and wide approaches may reveal findings of significance to both South African clinicians and to the greater research community.

2 PROBLEM STATEMENT

This project has the goal of investigating avenues for improved warfarin dosing in South African patients through the use of machine learning. To achieve this goal, two research aims are declared. The first is to determine whether training models on local data can lead to tangible benefits for South African clinicians and patients. The seconds is to compare the performance of "wide" and "deep" algorithms for warfarin dosing. In an effort to achieve these aims, the following research questions are proposed:

(1) Does training models on a South African dataset produce more accurate warfarin dose predictions for South Africans?
(2) How does the performance of "deep" algorithms compare to the performance of "wide" algorithms on local versus international datasets?
   (a) Is a bagged artificial neural network ensemble a viable machine learning technique for the application of automated warfarin dosing?
   (b) Does deep learning produce more accurate warfarin dosing than established neural network approaches?

Question 1 can be answered by training our own models on the PathCare dataset and comparing the clinically-relevant metrics [1] to those of a test subset (Section 3.2). Question 2 can be answered by comparing a deep neural network approach to an ensemble approach (both implemented using the same framework) on the same (PathCare) dataset using clinically-relevant metrics. Specifically, question 2a can be answered by comparing the predicted doses of the ensemble model to the clinically prescribed doses using clinically-relevant metrics, and question 2b can be answered by comparing the performance of a model created with a neural network of more than two hidden-layers to the performance of a neural network with two or fewer hidden layers, using clinically-relevant metrics.

[1] clinically-relevant metrics are detailed in Section 3.4.

3 PROCEDURES AND METHODS

3.1 Data Pre-Processing and Balancing

The successful training of a supervised machine learning model relies greatly on the quality of the data provided. There are two major criteria for determining the quality of a dataset for this purpose: whether it is *clean* and whether it is *balanced*. A clean dataset has the following properties:

- Every row has a value for every column
- A column corresponds only to a single variable
- Within a column, every value is of the same unit of measure

A balanced dataset is one which contains an equal proportion of the different classes of values that can occur in the domain of interest. It has been shown [34] that machine learning algorithms perform better on balanced datasets than on ones with a more "natural" distribution. If the data is unbalanced, the model may perform poorly in the cases on which it was inadequately trained.

3.1.1 Cleaning. The two datasets under consideration – the IWPC [36] and PathCare datasets – will need to be scrutinised for inconsistencies. Various pre-processing libraries exist (such as the one included in the *Keras* framework) which can be used to determine if the data fall within the above constraints.

3.1.2 Balancing. Within the PathCare dataset there are three natural classifications into which the data points fall: those for which the INR is greater than the therapeutic range (TR), within the TR, and below the TR. In these cases the warfarin dose should be decreased, left unchanged and increased accordingly. Should the data not be proportioned equally into these three classes, several courses of action are available. The two most commonly used methods are *oversampling* (in which the smaller classes are resampled until they approximate the size of the largest class) and *undersampling* (data points are removed from larger classes). The latter poses problems in that valuable data is lost; whereas the former leads to the possibility of overfitting [9]. The most appropriate technique to use will depend on the distribution of the dataset and may likely involve a combination of both approaches in order to retain the data's integrity.

3.2 Dataset Splitting Policy

Upon receiving the de-identified PathCare dataset, data pre-processing will take place. This will involve organising the data and dealing with missing values (see Section 3.1). Upon completion, between 5% and 20% (depending on the size of the dataset) will be randomly assigned to a *holdout* set. This will be separate from the active dataset used to train the models. The holdout set will only be used to perform the final evaluations and comparisons at the end of the project. Because the contents of this set is the same across all algorithms and team members, it makes for a fair evaluation of performance. The majority of the data will be used for training and validating new models. It is likely that the PathCare dataset will not be sufficiently large for a conventional train/validate split, which necessitates that cross validation (CV) will be required. Team members are free to use K-fold cross validation (KFCV) or leave one out cross validation (LOOCV) as is necessary and practical.

A final consideration is whether splitting should occur record-wise or subject-wise. The PathCare dataset contains multiple records for each patient. If the dataset is split record-wise, then the separate sets will be inherently linked by the shared patients from which the records are obtained. This would cause extremely biased results. To avoid this, the dataset will be split patient-wise. This may result in slight imbalances, but this can be easily remedied with standard techniques (Section 3.1).

3.3 Algorithm Development

The algorithms will be developed independently by the team members and compared at the end of the project. Development will be in the Python 3 programming language, with the use of a number of libraries, including *Keras* and/or *scikit-learn* for rapid prototyping of models, and *Tensorflow* for more intricate manipulation of neural networks.

Initially, simple linear regression (LR) models will be built on the PathCare dataset to establish a baseline for performance. Thereafter, simple artificial neural network (ANN) models will be built using the PathCare dataset to establish a benchmark for later models to beat. By evaluating more complex models against the LR *baseline* and ANN *benchmark* during the development stage, the team can have immediate and meaningful feedback on model performance.

Upon completion of the baseline and benchmark models, iterative development of the "deep" and "wide" algorithms can begin. In each iteration, a model will be developed by training on the dataset, then compared to the LR baseline and ANN benchmark using clinically-relevant metrics (see Section 3.4). Modifications can be made based on the results before the next iteration.

3.4 Evaluation

In order to accurately compare the performance of models, a set of clinically-relevant metrics are essential. The literature features a number of metrics for model evaluation, but a blend of statistically- and clinically-relevant metrics was used in the most notable studies. Section 5.5 elaborates on this topic. For the purposes of this project, three popular and reliable metrics will be utilised:

(1) MAE: Mean absolute error of dose estimates compared with actual doses.
(2) R^2 value: Closeness of fit between estimated and actual doses.
(3) PW20: Percentage of patients with dose estimates within 20% of the actual dose.

4 ETHICAL, PROFESSIONAL, AND LEGAL ISSUES

Because this research does not interact directly with patients, there are only very limited ethical concerns. These concerns revolve around the privacy and consent of the patients in the PathCare dataset. PathCare will be de-identifying the individuals in their dataset before it is employed for the project, rendering it impossible for anyone to link data points back to the patients from which they came. PathCare, as the owners of the dataset and the client in this project, are responsible for obtaining the consent of the patients via their own channels. Once received, the dataset will be held securely by the research team and will remain inaccessible to the public. An application for ethics clearance has already been submitted to the

Faculty of Science Research Ethics Committee, and is pending approval. In the event of publication of research findings, the project will follow the guidelines of the UCT *Authorship Practices Policy*, with all contributing parties being acknowledged for their work. Intellectual property directly derived from this project will be subject to UCT's *Intellectual Property Policy*.

5 RELATED WORK

5.1 Dosing with Artificial Neural Networks

After linear regression (LR), artificial neural networks (ANNs) are the next most common technique employed for warfarin dose prediction. Simpler ANNs sometimes perform less effectively than LR at some dosage levels on some ethnicities [20, 29]. However, in more sophisticated implementations, ANNs were superior to LR and other methods in their dose prediction accuracy [10, 37]. Additionally, it was found that a multi-layer perceptron (MLP) neutral network can outperform k-nearest neighbours (kNN) and model tree (MT) approaches, and is more responsive to ensemble techniques [10]. Zhou et al. attribute the success of ANNs to greater fault tolerance [37]. Despite widespread use of ANNs in many studies, no notable publications have been submitted on the topic of *deep learning* – using ANNs with many more hidden layers – for warfarin dosing. A recent article by Ching et al. [6] suggests this is due to the mismatch between individuals skilled in deep learning and individuals well-versed in biological and medical fields. This suggests that deep learning is an avenue worthy of investigation in the future.

5.2 Dosing with Ensemble Approaches

Ensemble machine learning methods involve synthesising the outputs from several, (often) simple models to obtain better performance than consulting individual ones [26]. There are numerous methods with which one can integrate the different base algorithms, however here we will only consider *bagging* which involves training multiple models of the same type with different samples of the dataset. Each base model is then given an equal vote to determine a final outcome. ANNs have been found to act as a particularly adept base model in conjunction with bagging [5].

Hu et al. [10] found that bagged ANNs outperformed a single ANN in terms of greater predictive accuracy and lower variance when applied to the warfarin dosing problem. ANN performance can vary highly depending on the correct choice of parameters for the model. They took the approach of iterating through a range of possible parameters to find a locally optimal configuration. Tao et al. [32] and Bashiri et al. [4] made use of evolutionary algorithms to select a population of models to form an ensemble and tune the parameters of an individual ANN, respectively. The success of the above approaches informs our decision to opt for a bagged ensemble model, using evolutionary algorithms to optimise the parameters of the base ANNs.

5.3 Pharmacogenetic and Clinical Data

All notable studies on warfarin dose prediction use one of two factor classes to develop models – either only clinical factors, or both clinical and pharmacogenetic factors. Common *clinical* factors include age, body mass, height, other medications the patient is taking, whether or not the patient smokes, and what other diseases the patient suffers from. *Pharmacogenetic* factors include all the genotypes for each of the SNPs (single nucleotide polymorphisms) associated with warfarin metabolism and interaction. The two most notable of these are found in *CYP2C9* and *VKORC1*, which explain approximately 40% of the individual variation in dose requirement [13]. Whilst many studies have reported improved model performance when using both clinical and pharmacogenetic factors [19, 24, 25, 27, 33], there is doubt as to whether pharmacogenetic dosing is actually clinically beneficial [2, 3, 10, 17]. Genetic testing is still very expensive, takes additional time and resources, and is unavailable in many parts of the world [10]. There is still no consensus about whether or not pharmacogenetic factors provide notable benefit to the model. More advanced machine learning approaches have produced significant gains in performance over previous models, despite many of them [10, 22, 28–30, 37] not incorporating any pharmacogenetic factors whatsoever. It appears that a better approach to improving warfarin prediction is found in making superior use of the data available. This suggests that even though the PathCare dataset does not contain any pharmacogenetic data, it may still allow us to develop local models that afford high precision at low financial and time costs.

5.4 Dataset Splitting

To prevent overfitting models to the data, it is essential to split the total dataset into multiple sets, train on one and test on the other. This is not only good experimental design, but an essential tool in training robust models of any kind. However, just as a model can overfit to training data, tweaking model parameters to increase performance can result in overfitting to the test data. This is the motivation behind holding back a small subset of the data to test the model at the very end of the experiment. Ideally, more data allows for more robust models through more stringent test splitting. In cases of small datasets (such as for warfarin dosing), it is tempting to reserve as much of the data as possible for training, but this makes evaluating the performance of the model difficult and could result in a model that does not generalise – which is disastrous in clinical practice. Studies of warfarin dose prediction have reported splits ranging from 50/50 [8, 11, 28] to 80/20 [19, 20, 29]. This is unsurprising given that the size of the cohorts in these studies ranged from a mere 174 [11] up to 4798 [20]. However, due to a lack of standardised terminology across different fields, it is difficult to interpret whether these reported splits included or excluded holdout sets. As a result, it is challenging to accurately evaluate the clinical significance of results in the literature. In recent years, many studies have begun using the International Warfarin Pharmacogenetics Consortium (IWPC) dataset [36]. Whilst this is advantageous due to its wealth of data, consistency, and size ($N \approx 5000$), it makes the likelihood of overfitting much greater. With many studies [19, 20, 28] competing to improve performance on this same dataset, the risk of (intentional or unintentional) overfitting is heightened. For this project, both holdout and cross-validation techniques will be used. Additionally, both the PathCare dataset (which is novel for this field) and the IWPC dataset [36] can be used to train and evaluate models. With more data available, stricter holdout and cross-validation policies can be implemented. The same initial holdout set will be

used to evaluate both team members' models, giving an accurate and unbiased comparison of performance.

5.5 Evaluation Methods

There have been a number of suggested approaches to evaluating model performance for warfarin dosing. Many standard statistical measures provide highly-precise measures of model performance, but do not take clinical relevance into account. The output variable is INR, which has a therapeutic range of 2.0-3.0 in most patients [1]. Using tighter target ranges for maintenance dosing does not achieve any improvement in anticoagulation control [23]. As a result, many studies use metrics that reflect this acceptable therapeutic range. The three most notable metrics for warfarin dose prediction will be used in this project (see Section 3.4). These afford both statistically- and clinically- relevant measures of model performance. They also allow for direct comparison with the reported performance of other studies.

(1) Mean absolute error (MAE) is widely used across warfarin prediction studies [8, 10, 20, 21, 28, 29, 31, 37].
(2) R^2 is the gold standard for statistical comparison and was utilised in numerous studies [7, 8, 11, 20].
(3) Percentage of patients with dose estimates within 20% of the actual dose (PW20) was used by a number of notable studies [8, 19, 20, 37].

6 ANTICIPATED OUTCOMES

6.1 Resulting System

The project should reach its conclusion with the development of a coherent software system for running and statistically evaluating the new machine learning models. This would likely take the form of a command-line interface for a *Python 3* program that can take in the name of the desired model and a vector of patient information as parameters, then return the model's prediction. It should also be able to compare all the models (including benchmarks and baselines) on any given dataset using clinically-relevant metrics (see Section 3.4), producing an easily-interpreted output. This output may take the form of a graphical rendering through the use of libraries such as *matplotlib*. All major design challenges will occur in the pursuit of this final system. A likely challenge is overwhelming training periods for the models, which may result from sub-optimal design choices in the development iterations. A development goal of this project is, therefore, to produce models that are not only accurate, but train rapidly on consumer hardware – enabling other members of the research community to easily replicate any results.

6.2 Expected Impact

If our hypotheses are correct, training models on local datasets will produce more accurate warfarin dosing suggestions for South African patients. If the resulting technology were to be implemented, it would improve the ability of clinicians to prescribe the correct warfarin doses to their patients, whilst minimising the time required to do so. It would also reduce the risk of severe haemorrhaging in patients and the number of return visits required to establish a therapeutic dose. If the proposed "deep" and "wide" algorithms reveal new insights into warfarin dosing in general, the greater medical community would benefit from the knowledge, which may lead to a future in which warfarin therapy is entirely safe and easily automated – freeing up valuable human resources to address other medical challenges. Even if the project finds no benefit in training local models or the new algorithms fail to outperform previous ones, the findings may reduce the problem space for future research and serve as evidence in the literature.

6.3 Key Success Factors

When evaluating the newly developed models, the clinically-relevant metrics (MAE, R^2, and PW20) will need to compare favourably with both those of the benchmark models and those reported in the literature. MAE values in the new models that are lower than those of the benchmark and literature would indicate success, whilst R^2 and PW20 values should be higher in the success case. If the clinically-relevant metrics are more favourable for models trained on the PathCare dataset than those trained on the IWPC dataset, then we may successfully conclude that training local models is valuable. If not, we have evidence to suggest that training local models may hold little value.

7 PROJECT PLAN

7.1 Risks

There are a number of core risks to the project. Aside from the risks common to all tasks of this nature – scope creep, missed deadlines, etc. – there are two primary classes of risk that are unique to this instance. The first class is associated with the PathCare dataset, which may be insufficiently large or lack critical parameters. In the worst scenarios, the dataset may be entirely unusable or may not even be provided. This class of risks is largely absorbed by the presence of the International Warfarin Pharmacogenetics Consortium (IWPC) dataset [36], which is the gold standard for warfarin dosing models. This freely available resource allows our project to supplement (or entirely replace) the PathCare dataset if necessary. In the scenario that the PathCare dataset lacks important attributes (such as ethnicity) for adequate training, random forest regression (RFR) models can be trained in place of artificial neural networks (ANNs). The results of the RFR methods can also be compared to results in the existing literature using the same metrics. The second class of risks is associated with failure to successfully implement specific machine learning algorithms. These risks are largely overcome by the wealth of libraries and resources available for developing simple, off-the-shelf models. The specifics of all risks are detailed in a risk matrix in the appendix.

7.2 Timeline and Milestones

Table 1 shows the most relevant internal and external project milestones. A full timeline of the project is detailed in the Gantt chart found in the appendix.

7.3 Resources

The most notable resource required is the PathCare dataset, which will be obtained directly from PathCare once ethical clearance has been obtained. Another notable resource is the IWPC dataset [36], which is open sourced and available for free download. The

Table 1: Project milestones, with critical milestones in bold.

Date	Description
2018-05-22	**Written project proposal submitted**
2018-05-28	**Proposal presentation**
2018-06-11	Revised project proposal submitted
2018-06-15	Project web presence goes live
2018-07-02	Data pre-processing and splitting complete
2018-07-09	Rudimentary models in testing
2018-07-27	**Feasibility demo to supervisor**
2018-08-06	Training completed, testing begins
2018-08-13	Testing completed, writing begins
2018-08-23	Draft papers given to readers
2018-09-03	**Final papers completed**
2018-09-04	Final code completed
2018-09-17	**Final project demonstrated**
2018-09-18	Project web page and poster completed
2018-10-03	Reflection papers submitted

development of models requires at least two modern computers with the open sourced *Python 3* programming language and freely available libraries such as *numpy*, *matplotlib*, *Tensorflow*, *scikit-learn*, and *keras*. Access to a high-performance computing cluster may be required if there is both sufficient need and justification for training on such architectures. The ongoing input of our supervisor will help guide development choices throughout the process, as facilitated by regular meetings.

7.4 Deliverables

The following are the expected deliverables for this project:

(1) Two advanced machine learning models that can output recommended warfarin dosage given a vector of patient information via a command-line interface.
(2) A series of statistical comparisons of the performance of these models against the LR baseline and simple ANN benchmark using clinically-relevant metrics (see Section 3.4).
(3) A series of statistical comparisons of the value of the PathCare dataset compared to the IWPC dataset [36] using the same clinically-relevant metrics.

7.5 Work Allocation

The initiation of the project has begun with both team members working in direct collaboration. Henceforth, the project will feature both independent and collaborative work. Team members will work together to clean, process, and split (see Sections 3.1 and 3.2) the PathCare dataset when it is received. Once this is completed and the holdout set is agreed upon, the project will fork into two branches. Gianluca will begin implementing "deep" algorithms (see Section 1), with a focus on deep learning techniques in artificial neural networks; whilst Neville will begin to develop "wide" algorithms (see Section 1), with a focus on ensemble approaches to artificial neural networks. Team members will develop baseline and benchmark models (see Section 3.3) independently to improve the robustness of the experimental design and ensure reliable results during the final evaluating stage. In that final stage, the team members will work together to thoroughly compare the performance of all models and the value of the PathCare dataset so as to address the research questions identified. Each avenue of inquiry has sufficient merit to stand as an independent project, but the final combination allows for more interesting and reliable comparisons.

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] Douglas C Anderson. 2009. Pharmacogenomics and warfarin. *American journal of health-system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists* 66, 2 (jan 2009), 121. https://doi.org/10.2146/ajhp080635

[3] 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

[4] M. Bashiri and A. Farshbaf Geranmayeh. 2011. Tuning the parameters of an artificial neural network using central composite design and genetic algorithm. *Scientia Iranica* 18, 6 (dec 2011), 1600–1608. https://doi.org/10.1016/j.scient.2011.08.031

[5] Leo Breiman. 1996. Bias, variance, and arcing classifiers. (1996).

[6] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-michael Agapow, Michael Zietz, Michael M Hoffman, Wei Xie, Gail L Rosen, Benjamin J Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M Cofer, Christopher A Lavender, Srinivas C Turaga, Amr M Alexandari, Zhiyong Lu, David J Harris, Dave Decaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K Wiley, Austin Huang, Anthony Gitter, and Casey S Greene. 2018. Opportunities and obstacles for deep learning in biology and medicine. *The Royal Society* (2018). https://doi.org/10.1098/rsif.2017.0387

[7] 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

[8] 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 (jan 2014), 29–37. https://doi.org/10.2217/pgs.13.212

[9] Haibo He and E.A. Garcia. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* 21, 9 (sep 2009), 1263–1284. https://doi.org/10.1109/tkde.2008.239

[10] 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

[11] 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 (mar 2014), 265–273. https://doi.org/10.1007/s00228-013-1617-2

[12] 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 arXiv:NIHMS150003

[13] 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

[14] 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 (sep 2009), 1482–1490. https://doi.org/10.1111/j.1538-7836.2009.03508.x

[15] 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

[16] 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

[17] Stephen E. Kimmel, Benjamin French, Scott E. Kasner, Julie A. Johnson, Jeffrey L. Anderson, Brian F. Gage, Yves D. Rosenberg, Charles S. Eby, Rosemary A. Madigan, Robert B. McBane, Sherif Z. Abdel-Rahman, Scott M. Stevens, Steven Yale, Emile R. Mohler, Margaret C. Fang, Vinay Shah, Richard B. Horenstein, Nita A. Limdi, James A.S. Muldowney, Jaspal Gujral, Patrice Delafontaine, Robert J. Desnick, Thomas L. Ortel, Henny H. Billett, Robert C. Pendleton, Nancy L. Geller, Jonathan L. Halperin, Samuel Z. Goldhaber, Michael D. Caldwell, Robert M. Califf, and Jonas H. Ellenberg. 2013. A Pharmacogenetic versus a Clinical Algorithm for Warfarin Dosing. *New England Journal of Medicine* 369, 24 (2013), 2283–2293. https://doi.org/10.1056/NEJMoa1310669 arXiv:NIHMS150003

[18] 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.

[19] B.F. Klein, T.E., Altman, R.B., Eriksson. 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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 (nov 2007), 1135–1141. https://doi.org/10.1007/s00228-007-0381-6

[25] 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

[26] R. Polikar. 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6, 3 (2006), 21–45. https://doi.org/10.1109/mcas.2006.1688199

[27] 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

[28] 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

[29] 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

[30] 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.

[31] 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

[32] Yanyun Tao, Yenming J Chen, Xiangyu Fu, Bin Jiang, and Yuzhen Zhang. 2018. Evolutionary ensemble learning algorithm to modeling of warfarin dose prediction for Chinese. 2194 (2018). https://doi.org/10.1109/JBHI.2018.2812165

[33] 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 arXiv:NIHMS150003

[34] Qiong Wei and Roland L. Dunbrack. 2013. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE* 8, 7 (jul 2013), e67863. https://doi.org/10.1371/journal.pone.0067863

[35] 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

[36] 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

[37] 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 (oct 2014), 1462–1464. https://doi.org/10.1016/j.ijcard.2014.08.062

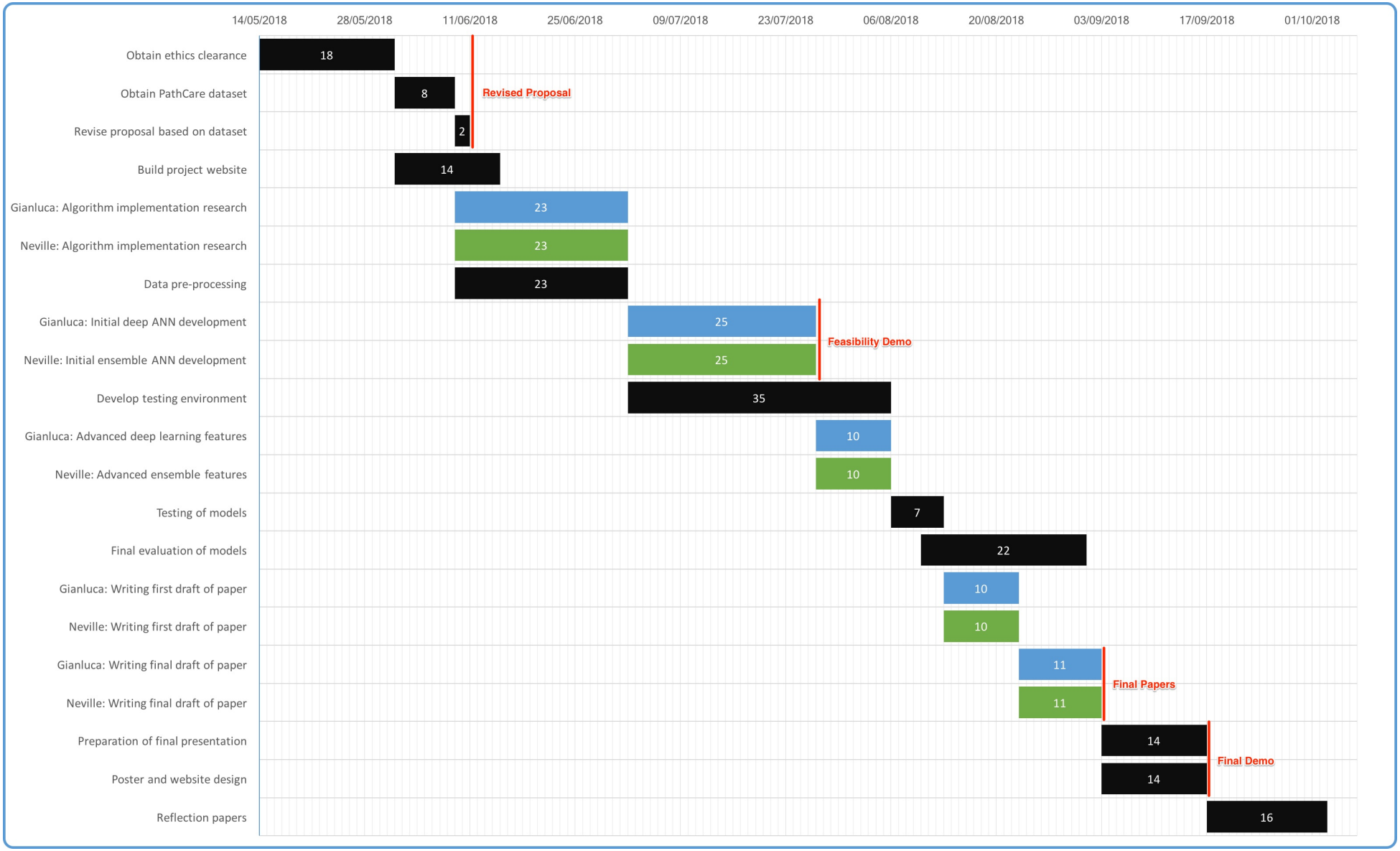

Figure A1: Gantt chart depicting prospective timeline and task durations (in days)

Table A1: Risk matrix for the monitoring, management, and mitigation of risks to the project

Risk	Probability 1-10	Impact 1-10	Consequence	Mitigation	Monitoring	Management
Scope creep / running over schedule.	7	9	Aspects of the project are not completed and the project as a whole is of lower value.	Constant adjustment of priorities, goals, and deadlines (accounting for planning fallacy).	Regular meetings with supervisor. Frequent communication between team members.	Remove non-essential aspects of project and re-focus on completing core goals.
The dataset proves inappropriate or insufficient for accurate training.	5	6	A core component of the project becomes unviable. Failure must be reported.	Make use of supplementary datasets (IWPC) and investigate alternative techniques.	Generate statistics of missing values and completeness before and after pre-processing. Frequent, direct comparison of model performance compared to linear regression on relevant metrics.	Supplement the dataset with the IWPC data or switch to the IWPC dataset completely. Switch to alternative techniques. e.g. Random forest regression (RFR)
PathCare fail to provide the dataset.	3	9	The core focus of the project is disrupted.	Make use of supplementary datasets, such as from the IWPC.	Regular contact with PathCare to ensure their commitment to provide the data.	Pivot to using the IWPC dataset and investigating novel machine learning approaches outside of local context.
One team member encounters unforeseen adversity and cannot complete their contribution.	3	7	The project is less cohesive and some research questions go unanswered.	Maintain resource-independence at all times. Keep some research questions for individuals.	Immediate communication of any change in circumstances to supervisor and team mate.	Narrow the research focus to the individual goals of the remaining team member and adjust timelines.
Fail to implement any models whatsoever.	1	10	The project is a total failure and there is nothing significant to report.	Use libraries (e.g. Keras) and standard patterns to quickly produce basic models, then tweak them.	Perform feasibility demo for supervisor early on, with at least one working model.	Use simpler algorithms and make use of libraries to simplify the implementation process.