A BAYESIAN MODELING AND ESTIMATION FRAMEWORK FOR PHARMACOGENOMICS DRIVEN WARFARIN DOSING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS INSTITUTE
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

SERDAR MURAT ÖZTANER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF MEDICAL INFORMATICS

JUNE 2014

A BAYESIAN MODELING AND ESTIMATION FRAMEWORK FOR PHARMACOGENOMICS DRIVEN WARFARIN DOSING

Submitted by **SERDAR MURAT ÖZTANER** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Medical Informatics, Middle East Technical University** by

Prof. Dr. Nazife Baykal,
Director, **Informatics Institute** _____

Assist.Prof. Dr. Yeşim Aydın Son,
Head of Department, **Health Informatics** _____

Assist.Prof. Dr. Tuğba Taşkaya Temizel
Supervisor, **Information Systems** _____

Prof. Dr. Remzi Erdem
Co-Supervisor, **Department of Pharmacology, Başkent University** _____

Examining Committee Members:

Prof. Dr. Ünal Erkan Mumcuoğlu
Medical Informatics, METU _____

Assist.Prof. Dr. Tuğba Taşkaya Temizel
Information Systems, METU _____

Assist.Prof. Dr. Aybar Can Acar
Medical Informatics, METU _____

Assist.Prof. Dr. Yeşim Aydın Son
Medical Informatics, METU _____

Prof. Dr. Ümit Yaşar
Department of Pharmacology, Hacettepe University _____

Date: 26 / 06 / 2014

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I fully cited and referenced all material that are not original to this work.

Name, Last Name: Serdar Murat Öztaner

Signature :

ABSTRACT

A BAYESIAN MODELING AND ESTIMATION FRAMEWORK
FOR PHARMACOGENOMICS DRIVEN WARFARING DOSING

Öztaner, Serdar Murat

Supervisor: Assist.Prof. Dr. Tuğba Taşkaya Temizel
Co-Supervisor: Prof. Dr. Remzi Erdem

June 2014, 80 pages

Recent studies have shown that the incorporation of genomics information into the drug dosing prediction formulations increases the accuracy of the drug dosing while decreasing the frequency of adverse drug effects. The current clinical approaches for drug dosing which are supported by the best pharmacogenomics algorithms explain only some percentage of the variance in dosing. The main objective of this study is to enhance the accuracy and efficacy of the warfarin dosing algorithms by using advanced methods of data mining and estimation. A novel framework based on Bayesian Structural Equation Modeling (SEM) is proposed for warfarin dosing. The proposed framework performs better than the state-of-the-art methods which make use of linear regression such Maximum Likelihood Estimation (MLE). The Bayesian SEM is a robust and effective approach for the estimation of warfarin dosing since it facilitates the exploration and identification of hidden relationships and provides the flexibility to utilize useful prior information for achieving better prediction results. Two independent data sets are used for comparison and validation purposes in this study: The combined multi-ethnic data set provided by the International Warfarin Pharmacogenetics Consortium (IWPC) and the Turkish data set. A series of data pre-processing techniques (feature selection, data imputation) are applied on both of the

data sets which contain common set of non-genetic features and genetic features including CYP2C9 and VKORC1 as the main pharmacogenomics variables. The non-linear model has converged with coefficients having small Monte-Carlo error and absolute values consistent with prior domain knowledge. The obtained pharmacogenomics warfarin dosing algorithm based on the non-linear Bayesian Structural Equation model accounts for up to 56.7% of the variation in warfarin dosage while the referenced pharmacogenomics warfarin dosing algorithms based on the linear regression model explains up to 51.2% of the variance. The prediction performances are also improved for both the data sets (47.4% and 51.7% respectively) compared to MLE (45.1% and 49.3%).

Key Words: Warfarin, Pharmacogenomics, Data Mining, Bayesian Structural Equation Modeling

ÖZ

FARMAKOGENOMİK VERİLERİ KULLANAN BAYES TEMELLİ WARFARIN DOZAJ MODELLEME VE TAHMİN ALTYAPISI

Öztaner, Serdar Murat

Tez Yöneticisi: Assist.Prof. Dr. Tuğba Taşkaya Temizel
Yardımcı Tez Yöneticisi: Prof. Dr. Remzi Erdem

Haziran 2014, 80 sayfa

Son zamanlarda yapılan çalışmalar, genomik bilgilerin ilaç doz tahmininde kullanılan formüllere dahil edilmesiyle bir yandan söz konusu ilaç dozlama formüllerinin başarı oranını artırdığını ve diğer yandan ilaç yan etkilerinin görülme sıklığını azaltığını göstermektedir. İlaç doz tahmininde kullanılan mevcut klinik yaklaşımlar en iyi farmakogenomik algoritmalarla desteklendiğinde bile ilaç doz varyasının ancak belirli bir yüzdesi açıklanabilmektedir. Bu çalışmanın ana amacı, gelişmiş veri madenciliği yöntemlerini ve tahmin algoritmalarını kullanarak warfarin dozlama algoritmalarının doğruluğunu ve etkinliğini arttırmaktır. Bu amaçla, hiyerarşik doğrusal olmayan karışım modeli (doğrusal olmayan karma etkiler modeli), Yapısal Eşitlik Modeli (SEM) kullanılarak kurulmuştur. Yapısal Eşitlik Modeli öncel bilgiden yararlanmak üzere Bayes yaklaşımı ile desteklenerek başta farmakogenomik faktörler olmak üzere diğer faktörlerin warfarin dozuna etkisini incelemek ve açıklamak amacıyla önerilmektedir. Çalışma kapsamında, veri ön-işleme teknikleri (özellik seçimi, eksik verinin tamamlanması vb.) Uluslararası Warfarin Farmakogenetik Konsorsiyumu (IWPC) tarafından sağlanan kombine veri seti üzerinde uygulanarak uygun bir veri kümesi sağlanmıştır. 5700 denekten elde

edilen veri kümesinde, aralarında ana farmakogenomik değişkenler olarak CYP2C9 ve VKORC1 yer alan 68 özellik bulunmaktadır. Doğrusal olmayan model, veri kümesi işlendikten sonra yakınsamış ve küçük Monte-Carlo hatalarına sahip ve öncel alan bilgisiyle tutarlı katsayı değerleri elde edilmiştir. Referans alınan çalışmada doğrusal regresyon modeline göre elde edilen farmakogenomik warfarin doz algoritması varyansı yaklaşık %51,2 oranına kadar açıklarken doğrusal olmayan Bayes Yapısal Eşitlik modeline göre elde edilen farmakogenomik warfarin doz algoritması varyanstaki değişkenliğine yaklaşık %56,7 oranında açıklama getirmektedir. Bayes tabanlı modellemenin kestirim başarısının da her iki veri kümesi için (%47,4 ve %51,7) Çoklu Doğrusal Kestirim yöntemine gore arttığı görülmüştür.

Anahtar Kelimeler: Warfarin, Farmakogenomik, Veri Madenciliği, Bayezyen Yapısal Eşitlik Modellemesi

DEDICATION

To my 16 months old daughter, DEFNE

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Assist. Prof. Dr. Tuğba Taşkaya Temizel for her continuous support of my Ph.D study. Without her persistent patience, motivation, guidance, and extensive knowledge this dissertation would not have been possible. She has helped me in all the time of research and writing of this thesis.

I would also like to thank to my co-advisor Prof. Dr. Remzi Erdem and committee member Prof. Dr. Ümit Yaşar, who introduced me to Pharmacology and Pharmacogenomics, and whose enthusiasm and immense knowledge for the "underlying structures" had lasting effect. I could not have completed my study without their encouragement, insightful comments, and valuable feedback on the domain.

Last but not the least, I would like to thank my family: especially my dearest wife Aslı Öztaner who share the burden with me, my little 6 months sweetie Defne Öztaner who has brought meaning and joy to my life, my father Önüt Öztaner and my mother Nesrin Öztaner who gave birth to me at the first place and have supported me throughout my life, my brother Burak Öztaner and my sister Şebnem Öztaner who have been by my site with their understanding and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

An important research challenge for patient safety is to minimize Adverse Drug Reactions (ADRs) [1]. Drugs are never completely safe and in addition to their desired effect, they also cause side effects. ADRs are among the prominent causes of hospitalization and death [2]. Although drugs are developed and approved after a well-defined, complicated and firmly controlled process, pharmacovigilance (post-marketing drug analysis) through SSSs (spontaneous surveillance systems) constitutes a very import role in detecting adverse drug effects once the drugs are marketed [3], [4], [5].

It is not a simple task to decide whether the response of patient to a drug will be good or bad –or even there is no response at all-, since drugs are produced and marketed according to a "one size fits all" system. Although the tendency to ADRs might emanate from both non-genetic and factors, it is argued that genetics has a substantial role in drug responses. Indeed, it has been indicated by various studies that variations in drug responses are associated with genetic markers as well as non-genetic factors such as age, gender, and ethnicity [6].

The emerging field of Pharmacogenomics (PGx), as a potential application of personalized medicine [7], is the study of how and why the personal genetic differences alter drug response [8], [9], [10]. PGx has been increasingly popular and important due to the advances in pharmacology and human genomics. Recent studies have shown that the incorporation of genomics information into the drug dosing prediction formulations increases the accuracy of the drug dosing while decreasing the frequency of adverse drug effects [6], [8], [11], [12], [13].

The main objective of this study is to enhance the accuracy and efficacy of the drug dosing algorithms by using advanced methods of data mining and estimation. Therefore, the primary motivation is to build a generic and robust drug warfarin dosing estimation model which is supported by pharmacogenomics data and based on Bayesian Structural Equation Modeling [14], [15]. This estimation of model is expected to provide a computational framework for the practitioners and pharmacology experts so as to assist them in their clinical and scientific studies. The current clinical approaches for drug dosing which are supported by the best pharmacogenomics algorithms explain only some percentage of the variance in dosing.

Due to the non-linear relationships between dose and response, a non-linear model is

developed by using Structural Equation Modeling (SEM) [16], [17], [18], [19]. A Bayesian approach to SEM is proposed in order to study and explain the effects of pharmacogenomics factors on drug dosing along with the other factors (i.e. flexibility to incorporate the useful prior information into the model). Bayesian model fitting usually depends on Markov Chain Monte Carlo (MCMC), which involves simulating draws from the joint posterior distribution of the model unknowns through a computationally intensive procedure [14], [20].

Warfarin, a commonly used anti-coagulant drug, is chosen for the study as it has a "narrow therapeutic window"[1]. The under-dosing and over-dosing of warfarin can be disastrous due to the thrombotic or hemorrhagic adverse reactions respectively. Since the initiation of warfarin therapy based on clinical procedures is risky and problematic, various studies have suggested warfarin dosing algorithms [11], [12], [21], [22], [23], [24].

There is substantial individual variation in warfarin response and 17−25% of this variability can be accounted for the clinical factors alone [13], [22], [25], [26], [27]. It was stated that 52−54% [22], [23], [25], [27] or about 55% [11] of the variability could be explained for some ethnic groups by including the polymorphisms of genes coding the enzymes CYP2C9 and VKORC1 and using various pharmacogenomics algorithms based on stepwise linear regression [11], [12], [22], [25]. Most of these pharmacogenomics algorithms made use of both non-genetic factors (BSA, age, race, target-INR (International Normalized Ratio), amiodarone in-take, smoking status, DVT (Deep Vein Thrombosis) or PE (Pulmonary Embolism)) [26] and well-known warfarin related Single Nucleotide Polymorphisms (SNPs) such as CYP2C9*2 (rs1799853), CYP2C*3 (rs1057910), VKORC1:-1639G>A (rs9923231) and VKORC1: 1173C>T (rs9934438) [27] [28].

In addition to the polymorphisms of CYP2C9 and VKORC1, CYP4F2 is also related with the variations in warfarin dosing. Recent studies have reported that polymorphisms in CYP4F2 might also account for up-to 7% of the variation [31] [32]. Another study by Özer et al. [21] on a group of Turkish patients deduced a linear regression model including CYP2C9*2, CYP2C9*3, VKORC1 −1639, and CYP4F2 (rs2108622) variants, as well as age, which explained 39.3% of the overall inter-individual differences in the warfarin dose.

This study aims to improve the accuracy and efficacy of the warfarin dosing by using advanced methods of data mining and data estimation. A novel framework based on Bayesian Structural Equation Modeling (SEM) is proposed for warfarin dosing. The proposed framework performs better than the state-of-the-art methods which make use of linear regression such as Maximum Likelihood Estimation (MLE). Additional genetic (VKORC1:1173C>T, CYP4F2) and non-genetic factors (concomitant medication, co-morbidities) are also incorporated into the proposed model to improve the accuracy of pharmacogenomics driven prediction.

Two independent data sets are used in this study: 1) The combined multi-ethnic data

[1] Therapeutic window is an indicator for estimation of drug dosage which can cure disease effectively without going beyond the safety range.

set provided by the International Warfarin Pharmacogenetics Consortium (IWPC) [11], [29], [30], [31]; and 2) The Turkish data set [21]. The IWPC data set and the Turkish data set contain common non-genetic and genetic features. Thus, the model for communized features can be validated by applying on both of the data sets. A series of data pre-processing techniques (feature selection, data imputation) are applied on the data sets so as to achieve better results from the Bayesian estimation.

Apart from the polymorphisms of the CYP2C9, VKORC1 and CYP4F2 genes [32], clinical factors have been accounted for the variability in warfarin dose requirements, including age, race, weight, height, gender, race, smoking status and medications (Taking an Enzyme Inducer, Taking an Enzyme Inhibitor or Substrate) [23], [33], [34], [35], [36]. Dietary components such as vitamin K intake or consumption of tea and coffee (or alcohol) may also influence warfarin dose requirements [37], [38]. Thus, new factors are introduced to the model. For instance, Rifampin as an enzyme inducer and Fluconazole, Fluvastatin and Lovastatin (in addition to Amiodarone which is also used in the reference studies) as enzyme inhibitors are included in the model. Aspirin is also introduced as an intake similar to smoking. By this, the secondary motivation of the study which is introducing new factors into the model is achieved. There are some other medications such as Simvastatin, Atorvastatin, Carbamazepine and etc. which are indicated in the IWPC data set, but these medications are not shown to be accountable for the variability in warfarin dosing due to the lack of data.

IBM® SPSS® AMOS™ (Analysis of Moment Structures) v21 software **[39]** was utilized for the application of Bayesian SEM analysis and inference based on MCMC. Informative (non-diffuse) priors were used for both of the IWPC and Turkish models so as to avoid negative variance estimates and other improper estimates since good prior information have been provided for some attributes by earlier studies. Based on the expert knowledge and the prior literature, several variants of models are suggested. Among these, the model having the best fit is selected based on the appropriate convergence diagnostics. The obtained pharmacogenomics warfarin dosing algorithm based on the non-linear Bayesian Structural Equation model explains up to 56.7% of the variance in warfarin dosage while the MLE explains up to 51.2% of the variance.

CHAPTER 2

LITERATURE REVIEW

This chapter puts forth the previous work to establish the patient safety by minimizing the ADRs of drugs via pharmacovigiliance and pharmacogenomics studies [9]. Section 2.1 introduces the concept of personalized medicine with regards to ADRs. Section 2.2 explains the Pharmacovigilance (PV) which is concerned with defining the outcomes of medical products that human life and with mitigating the associated risks of these outcomes. Section 2.3, on the other hand, expresses the emerging field of Pharmacogenomics (PGx) which analyses how genes can affect a person's response to drugs. Section 2.4 describes the factors that are accountable for the variations in warfarin dose and Section 2.5 mentions the previous studies that have been carried out to predict warfarin dose by utilizing these factors. Sections 2.6 and 2.7 express the statistical techniques that are made use of during the course of this study.

2.1 Personalized Medicine

Any health care service is a set of complex processes directly affecting human beings and has a level of innate unsafety at every point of each process. Adverse reactions may arise from problems in practice, procedures or systems. Overarching system-wide effort has to be exerted to minimize these adverse events and establish the maximum patient safety which is the main objective of health care. One of the research priorities for patient safety is to minimize adverse events caused by drugs [1]. Drugs are also never completely safe and in addition to their desired effect, they also cause side effects. For a drug to be, and remain, on the market its benefit must outweigh its perceived risk and it must be therapeutic in an acceptably high proportion of treated individuals.

WHO[2] defines ADR (Adverse Drug Reactions) as "A response to a drug which is noxious and unintended, and which occurs at doses normally used in man for the prophylaxis, diagnosis, or therapy of disease, or for the modification of physiological function" [2]. As reported in the widely known book “To Err Is Human” published by IOM[3], preventable adverse reactions are placed in the top 10 foremost causes of death and give rise to health care costs of $3.6 billion per year in the United States as

[2] World Health Organization (http://www.who.int)
[3] Institute of Medicine (http://www.iom.edu)

of 1997 [1]. Moreover, AMA[4] states that just around 50% of patients react sufficiently to drugs, and in fact, many prevailing medications are correlated with a significant risk of treatment failure or drug toxicity [40] [41]. The results of the meta-analysis of studies from 1966-1996 in the United States indicates that the total incidence of serious ADRs is 6.7%, of which 4.7% are accountable for admission and 2.1% occurs after admission, with an overall mortality rate of 0.32% [41]. A more recent study including 125 in-patients shows that 19% of the patients, who suffer from ADRs, spend 6.5 days longer in hospital than those without ADRs [40].

Drugs are extensively tested using animals and via clinical trials in humans, before they are marketed. These tests express much about the efficacy of drugs but delineate relatively little about the safety of drugs due to the following limitations of most clinical trials [42] [43]:

- Homogeneous populations: Most clinical trials exclude specific groups such as children, pregnant women, and people who are old or have chronic diseases by just covering comparably healthy people with only one disease.
- Sample size: Small sample size (up to 1000 patients) decreases the likelihood of finding rare adverse effects.
- Limited duration: Trials that last for a short period of time inhibit the long term consequences such as cancer.
- Inability to predict the real world: Clinical trials are carried out for one drug while patients often use more than one drug concurrently in the real world. Therefore, the drug interactions which can be significant in the real world cannot be encountered during clinical trials.

Since in the pre-marketing period, not all of the adverse reactions can be detected, the monitoring of drug effectiveness and safety must continue after marketing [44]. Ongoing post marketing surveillance is needed for all drugs on the market, not just new drugs, as evidence of increased risks (as well as decreased risks) continues to be found through the lifetime of drugs, as for the example of aspirin and increased risk of Reye syndrome.

As drugs are produced according to the *average* patient, it is not easy to determine whether a person will respond well, badly, or not at all to a medication. Personalized medicine is a medical approach which addresses to this challenge by proposing medications being customized for the individual patient. In the context of the personalized medicine, the customization of healthcare does not only relate to the use of medications but also the medical decisions, practices, and applications. Note that, the use of genetic information plays a major role in certain forms of personalized medicine [45]. The emerging field of PGx as a potential application of personalized medicine studies how personal genetic discrepancies affect drug response [6], [8], [9].

[4] American Medical Association (http://www.ama-assn.org/)

2.2 Pharmacovigilance

Pharmacovigilance (PV) which is also referred to as "the post-marketing drug safety" is the pharmacological science related to the detection, assessment, understanding and prevention of adverse effects [44], [46]. Since substantial amount of ADRs are observed in post-marketing period, efforts have been made to enhance the recognition of ADRs by mining the large databases generated by spontaneous surveillance systems (SSSs). SSSs which are also called as spontaneous reporting systems (SRSs) rely on the principle that when there are great numbers of reports, matching the proportion of reports of an adverse reaction with similar drugs may signal the rare and hidden patterns. There are various SSSs:

- The Adverse Event Reporting System (AERS) of FDA: AERS[5] receive the event reports directly or in batches periodically. AERS contains over four million reports of adverse events from 1969 to the present. Curated data is provided publicly on quarterly basis.
- The Canadian Adverse Drug Reaction Information System (CADRIS) of Health Canada: CADRIS[6] includes over 160,000 suspected AR reports that have been prepared in Canada since 1965.
- EudraVigilance[7] (European Union Drug Regulating Authorities Pharmacovigilance) is the European data processing network and management system for reporting and evaluating the suspected adverse reactions during the development of new drugs. It also follows the marketing authorization of medical products in the European Economic Area (EEA).
- The VigiBase[8] is the most comprehensive and largest data resource in the world, and is built and maintained by the Uppsala Monitoring Centre (the UMC) on behalf of the World Health Organization. VigiBase™ contains more than 3.8 million case reports from 82 nations. Around 50,000 new reports are added to it on quarterly basis. All of these cases can be accessed to by health professionals.

These surveillance systems are responsible for the early detection of serious ADRs and the majority of drug withdrawals from the market such as the withdrawal of Vioxx (rofecoxib) [47]. To extract knowledge from the pharmacovigilance databases generated by SSSs, statistical data mining methods ("association finding", "disproportionality analysis") are utilized within the domain of Knowledge Discovery in Databases (KDD).

As stated in [48], "Knowledge Discovery is defined as the nontrivial extraction of hidden, previously unrecognized, and likely useful information from data. KDD is frequently utilized in a wide range of pattern finding practices and profiling, such as surveillance, fraud detection, marketing and scientific discovery; moreover, it constitutes the core of pharmacovigilance applications [49] [50]. KDD is an overall

[5]AERS:http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm
[6] CADRIS: http://www.hc-sc.gc.ca/dhp-mps/pubs/medeff/_fs-if/2005-cadris-2/index-eng.php
[7] EUDRAVIGILANCE:http://eudravigilance.emea.europa.eu
[8] VIGIBASE: http://www.umc-products.com/DynPage.aspx?id=4910&mn=1107

process which incorporates several steps to derive knowledge from data. These steps are data preparation, data selection, data cleaning, incorporation of appropriate prior information, and proper interpretation of mining results. Figure 1 illustrates a more detailed approach for the KDD Process. Note that, the data mining is actually a step in the KDD process that comprises the application of data analysis and discovery algorithms over the data to generate a particular list of patterns [48]. The data-mining relies massively on known techniques from machine learning, pattern recognition, and statistics to extract patterns from data.

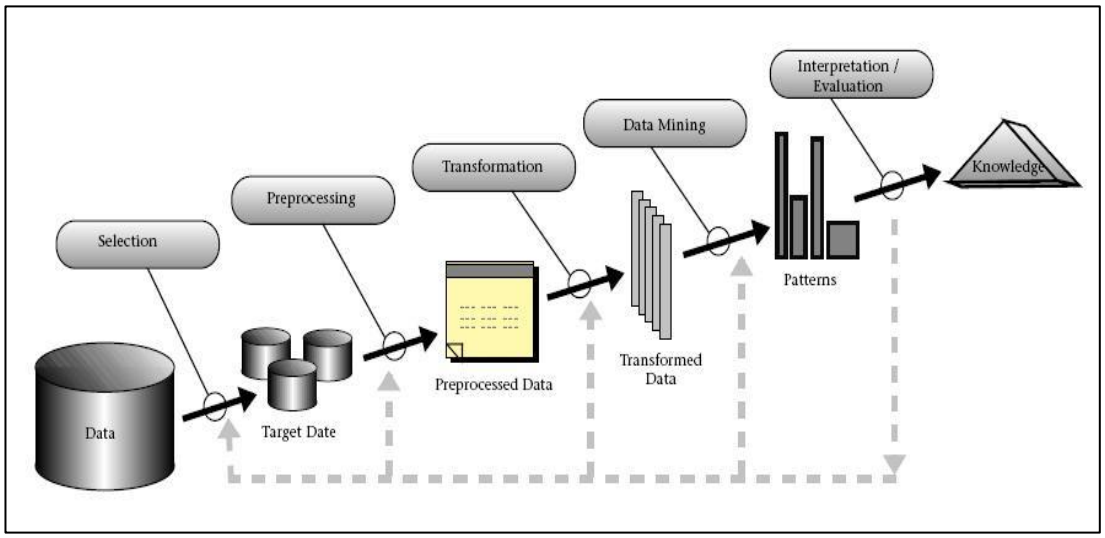

Figure 1: Steps of the KDD Process [48]

Although the reasoning and the methodology of the various quantitative data mining approaches differ, they all try to set forth to which extent the number of observed cases deviates from the number of expected cases. The MGPS (the Multi-item Gamma Poisson Shrinker) algorithm of FDA explores the ratio of an observed ADR to the total number of ADRs in order to detect a signal using Bayesian statistical analysis [51]. WHO UMC has also developed a system called as "Bayesian Confidence Propagation Neural Network" (BCPNN), which uses a feed-forward propagation neural network for searching drug-effect and drug-drug interactions by applying Bayesian statistics and Information Theory to quantify the unexpected signals. BCPNN can handle large amount of data containing incomplete data and be used with complex variables [52]. Figure 2 illustrates the overall surveillance of KD process in the drug safety databases.

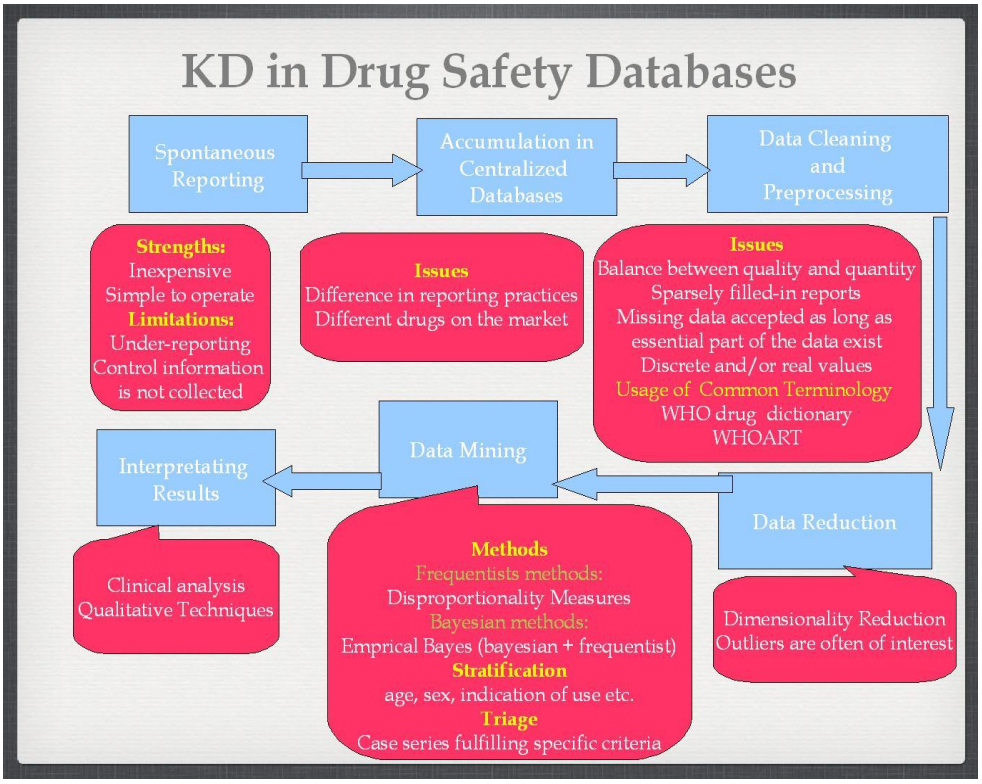

Figure 2: Knowledge Discovery in Drug Safety Databases

Although the knowledge discovery efforts have been successful in identifying new drug-ADR interactions, owing to the KDD approaches, there are problems due to the limitations of SSSs:

1. The reports have been collected primarily on a voluntary basis. Therefore, the surveillance is passive and there are often missing and incomplete data in reports.
2. There are many factors that affect the number of reports received such as under and duplicated reporting of adverse reactions within both voluntary and mandatory spontaneous surveillance systems.
3. Reporting may be subjective because ADR reports are issued by the opinion or observation of the individual reporter based on suspicious associations.
4. There may be FP (false positive) cases often revealed by the inclusion of a particular reaction which are not actually caused by the suspected drug.
5. Drug-drug interactions often make the reporting complex. It may not be always possible to determine which of the drugs are responsible from the reaction [52], [53].

There have been studies to improve the surveillance reporting mechanisms by establishing active surveillance via standardized and collaborative systems, and also to enhance the data mining techniques used [1]. A retrospective evaluation of the WHO BCPNN demonstrates that BCPNN provides a good overall sensitivity rate along with a rather lower specificity rate [54]. The performance of the method is also limited by the fact that it does not consider the semantic information to encode the adverse events in case reports [55]. Thus, efforts continuously have been being

exerted to improve the performance of BCPNN method in the new versions of Vigibase system.

2.3 Pharmacogenomics

Although the ADRs may be induced by both non-genetic and genetic factors, it is argued that genetics has an important role in drug responses. Various studies indicate that the variations in drug responses can be correlated with genetic markers as well as non-genetic factors such as age, gender, and ethnicity [6], [9], [56].

For instance, there is an increased risk of severe ADRs in children due to the following reasons [57]:

- 75% of approved drugs used in children are untested in pediatric populations
- Young children cannot evaluate or express their own response to medications
- Pediatric dosage forms not available
- Children metabolize drugs differently than adults.

Therefore, it is vital to develop genotype-based dosing guidelines to ensure the drug safety and avoid severe ADRs [26], [58]. PGx is the branch of pharmacology, which studies the effects of genetic variations on drug response by correlating gene expression or SNP with efficacy or toxicity of a drug. PGx focuses on pharmacogenes, where a pharmacogene is any gene involved in the response to a drug. As of 21 May 2014 (NCBI dbSNP[9] Build 141), there are almost 44 million validated SNPs for human beings. The number of SNPs that have been submitted to dbSNP is actually over 260 million and the number is growing on day to day basis. When the HapMap[10] (haplotype mapping of the human genome) Project is completed, there will be enormous amount of genomics data to be utilized for analysis and identification of the risk haplotypes for specific drugs or drug categories. These data and information supported by PGx studies have been accumulated in various data repositories and knowledge bases. These publicly data repositories and knowledge bases are used in the study of serious ADRs in the context of PV and PGx. Figure 3 depicts an example of how these repositories are linked based on a widely-known pharmacogene, thiopurine methyltransferase (TPMT) [59].

[9] http://www.ncbi.nlm.nih.gov/SNP/index.html
[10] http://www.hapmap.org/

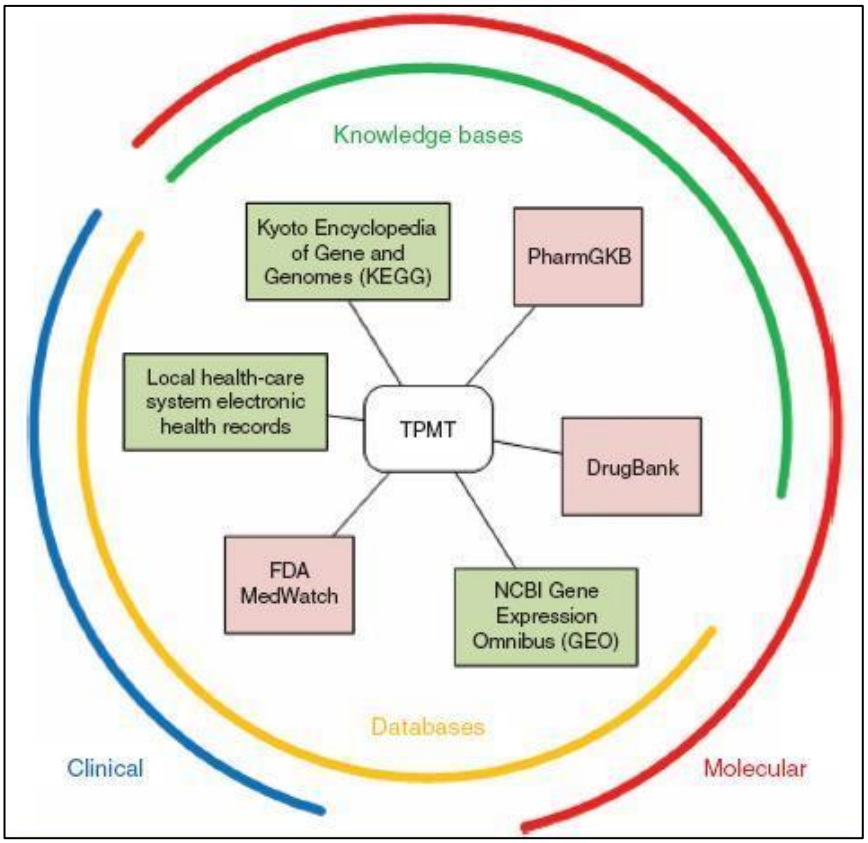

Figure 3: Integrating Life Sciences Databases for better pharmacovigilance [59]

TPMT exists in several publicly accessible information sources. The Pharmacogenomics Knowledge Base (PharmGKB), the Kyoto Encyclopedia of Genes and Genomes (KEGG), and the DrugBank contain knowledge on TPMT, while databases such as clinical electronic health records, MedWatch, and GEO provide observed measurements or characteristics for TPMT.

PGx can be used to improve the effectiveness and efficiency of health care safety by maximizing the probability of desired outcomes and minimizing the risk of ADRs making use of a person's genetic features. PGx can also be used to reduce the number of post-approval drug withdrawals [45]. Moreover, PGx may improve the quality of health care for chronic diseases. The current treatment approach for these diseases is to slow down their progression and lessen their symptoms. PGx may support the therapeutic approaches to clinical diseases by more effectively abolishing the symptoms, reducing the health care costs, and avoiding ADRs at the same time [45], [60].

The advances in the fields of pharmacology, genetics, and human genomics have caused PGx to emerge as a new promising field of study. However, the report "Realizing the Potential of Pharmacogenomics: Opportunities and Challenges" by the Department of Health and Human Services of USA indicates that the benefits of PGx on a large scale will be realized in the long-term [60]. There are few practical applications of PGx because the current health insurance policies discard the recompense, of most screening test, restrain PGx innovation by discouraging the

usage of PGx tests and therapies by health care providers.

Additionally, the infrastructure of current health information technology cannot meet the needs for researching PGx technologies and supporting PGx applications. In conclusion, PGx has the potential to provide better targeted and more effective treatments for patients while reducing the adverse effects due to the advances in pharmacology, pharmacogenomics and information technologies but there are several practical and clinical obstacles to be avoided.

2.4 Factors of Warfarin dosing

There are several reasons to test the hypothesis that pharmacogenomics can help to reduce drug toxicity by using warfarin:

- It is commonly used
- It has a narrow therapeutic/toxic ratio
- The under-dosing and over-dosing of warfarin can be disastrous
- It is influenced by widely-known genetic polymorphisms
- There is a substantial personal variation in warfarin response

Incidents associated with warfarin have placed it in the "top 10 drugs" for ADR related hospitalizations in the US [61], [62]. Between 2007 and 2009 warfarin accounted for 33% of drug-related hospitalizations for adverse events in the US [62]. Therefore, it is crucial to achieve a safe and effective level of anticoagulation for patients starting warfarin.

As stated by [13], [22], [23], [25], [26], [27], 17−25% of the variability in the therapeutic warfarin dose can be associated with the clinical factors alone. Figure 2 roughly depicts the variability in warfarin dose based on the major factors. It is also shown that 52−54% [22], [23], [25], [27] or about 55% [11], [63] of the variability in dosing can be explained for some ethnic groups by including the polymorphisms of genes coding the enzymes CYP2C9 and VKORC1 and using various pharmacogenomics algorithms based on stepwise linear regression. Most of these pharmacogenomics algorithms made use of both non-genetic factors (BSA, age, race, target-INR (International Normalized Ratio), Amiodarone in-take, smoking status, DVT (Deep Vein Thrombosis) or PE (Pulmonary Embolism)) [26], [31], [33], [34] and well-known warfarin related SNPs such as CYP2C9*2 (rs1799853), CYP2C*3 (rs1057910), VKORC1:-1639G>A (rs9923231) and VKORC1: 1173C>T (rs9934438) [26], [28], [35], [64], [65], [66].

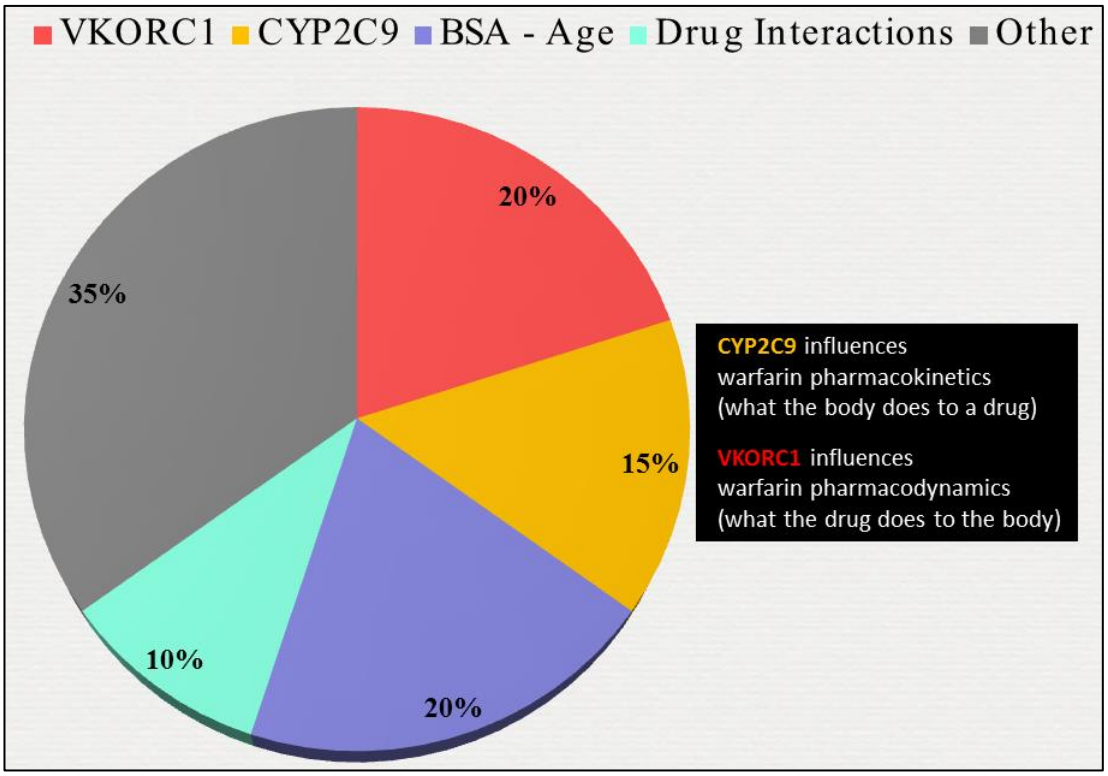

Figure 3: Major Factors of Warfarin Dosing

The SNPs of CYP2C9 and VKORC1 account for ≈ 40% (≈ 15% and 25% respectively) of the variation in warfarin dose [67], [68]. CYP2C9 is a member of the family called as Cytochromes P450 (CYPs) which belongs to the superfamily of proteins functioning in enzymatic reactions. The drug interaction table for P450 [64] shows all the pharmacogenetic influences and the corresponding drugs in terms of substrates, inhibitors and inducers. The most effective variants of CYP2C9 are the *2 (rs1799853) and *3 (rs1057910) polymorphisms which are associated with lower dose of warfarin [26], [35], [36], [69], [70]. The daily maintenance dose has to be decreased 15-20% for the patients who carry the mutations of CYP2C9*2. Similarly, the patients who carry the mutations of CYP2C9*3 require 30-40% lower daily maintenance dose. This fact also indicates that CYP2C9*3 plays a more significant role compared to CYP2C9*2 in explaining variations of warfarin dose [31], [36] .

CYP2C9*2 are more commonly observed in the Caucasian population (≈8-19%), but much less frequently in African-American and African populations (≈ 2-4%). It is not seen in Asian populations (0-0.1%) [69], [66], [65]. CYP2C9*3 has a similar frequency distribution in Caucasian (≈3-16%), African-American/African (≈0-2.5%) and Asian (≈1.1-3.6%) populations [10], [31], [66].

VKORC1 (vitamin K epoxide reductase subunit 1) converts vitamin K epoxide to vitamin K which is needed by certain proteins as a cofactor for coagulation. Warfarin inhibits VKORC1 which reduces the amount of vitamin K. For that reason, variations of VKORC1 are very effective for warfarin dosing. Several studies have shown that SNPs of VKORC1:-1639G>A and VKORC1: 1173C>T or C6484T (rs9934438) are accounted for significantly lower warfarin doses [26], [28], [31],

[35], [67], [68], [71], [72]. Since VKORC1: 1173C>T is indicated to be in "near perfect linkage disequilibrium" with VKORC1:-1639G>A, these two variants are almost equally predictive for warfarin dosing [67].

Individuals who carry the A allele in VKORC1:-1639G>A haplotypes require a lower initial dose of warfarin than those who carry the G allele. This is increased per A allele. That is, heterozygous carriers of the A allele respond to an intermediate warfarin dose, whereas homozygous carriers of the A allele respond to the lowest dose of warfarin. Homozygous carriers have the highest risk for warfarin-related adverse reactions. Almost 28% decrease in the therapeutic warfarin dose per A allele is required, and this SNP is the one of the most important predictors for the initiation dose for warfarin, [67], [68], [71], [73], [74].

The frequency of VKORC1:-1639G>A varies according to ethnic groups. It is actually the majority allele (around 90%) in Asian populations, and is also quite common in Caucasians having an allele frequency around 40% [31] [12] [74].

It has been indicated that the variant of CYP4F2 (rs2108622) accounts for warfarin related enzyme activity (4-12% increase in warfarin dose per T allele) [32], [72], [75] [76]. A recent study on Turkish population also demonstrates that CYP4F2 is associated with 2.8% increase in warfarin dose per T allele [21]. There is also another variant namely VKORC1: 3730G>A or G9041A (rs7294) which might be accounted for a higher warfarin dose but this varıant is not included in this study [31], [67].

The results of the pharmacogenomics studies have stimulated the U.S. Food and Drug Administration (FDA) to update the warfarin drug label to include information about genetic polymorphisms affecting dosing in August 2007 [13]. The widely annotated genetic variations are tabulated in Table 1.

Table 1: Widely known SNPs of CYP2C9 and VKORC1

Variant Name: **CYP2C9*2** Disposition: **rs1799853 at chr10:96692037** in CYP2C9 This SNP has been indicated to affect the clearance of various other drugs (fluvastatin, glipizide, phenytoin, tolbutamide) as well as warfarin.
CYP2C9*3 **rs1057910 at chr10:96731043** in CYP2C9 This SNP has been indicated to affect significantly the clearance of various other drugs (fluvastatin, glipizide, phenytoin, tolbutamide) as well as warfarin.
VKORC1:G9041A; VKORC1:3730G>A **rs7294 at chr16:31009822** in VKORC1

This SNP might be linked with a higher warfarin dose.
Variant Name: **VKORC1:G3673A; VKORC1:-1639G>A** Disposition: **rs9923231 at chr16:31015190** in VKORC1 This SNP is significantly associated with a lower warfarin dose based on both in vivo and in vitro evidence.
VKORC1:C6484T; VKORC1:1173C>T **rs9934438 at chr16:31012379** in VKORC1 This SNP is also significantly associated with a lower warfarin dose and is said to be in near perfect linkage disequilibrium with VKORC1:-1639G>A.

Apart from the SNPs of CYP2C9 and VKORC1, several other polymorphisms (such as rs2108622 at chr19:15851431 in CYP24F, rs2292566 at chr1:224086276 in EPHX1, rs339097 at chr7:128186460 in CALU), genetic, and clinical factors have been accounted for the variability in maintenance warfarin dose, including age, race, weight, height, gender, race, smoking status and medications (Taking Enzyme Inducer, Taking Amiodarone) [26], [77], [56], [78].

Dietary factors such as vitamin K intake, food supplements, and alcohol consumption may influence warfarin dose requirements. Warfarin works by inhibiting the recycling of fat-soluble vitamin K, thus higher consumption of fat-soluble vitamin K (found in green vegetables) can theoretically reduce the action of warfarin. In addition, low or inconsistent vitamin K intake might contribute to variations in anticoagulation control [26], [63].

Numerous medications can affect warfarin dose. Patients using amiodarone were estimated to require 29% lower doses in a study of 369 adults on stable warfarin therapy [34], [37], [63].

Dose requirements decrease with age because of increased responsiveness and/or decreased clearance. A study conducted on 297 patients on stable warfarin doses for the ages between 20 to 90 years demonstrated that the mean of daily dose requirements for warfarin decreased by 0.5 to 0.7 mg per decade [27], [33], [79].

Various illnesses can affect warfarin dose requirements. Patients with liver disease, malnutrition, de-compensated heart failure, hypermetabolic states (e.g. febrile illnesses, hyperthyroidism), hypertension, renal insufficiency, malignancy and different therapeutic indications for warfarin (e.g. prosthetic heart valve, current venous thromboembolism) have been indicated to alter dose requirements [23], [33].

2.5 Reference Studies

The previous studies which proposed pharmacogenomics driven formulations for the prediction of warfarin dosage used a model based on linear regression [11], [12], [21], [22], [23], [24]. Roper et al. [25] carried out a comparative study to validate three published warfarin dosing algorithms by Sconce et al. [12], Anderson et al. [24], IWPC and Klein et al. [11] and the WarfarinDosing website [80] based on the studies of Gage et al. [22], [23]. The predicted dose requirements were compared with the actual maintenance dose for each patient within the therapeutic INR of 2.0 to 3.0 [11], [21], [22]. It was concluded that all linear regression based methods produced similar results explaining only %37.7 to %45.8 of the variance in warfarin dosage in the IWPC training data set as indicated in Table 2. Note that, all the algorithms tested in the scope of the study by Roper et al. [25], are mainly based on CYP2C9*2, CYP2C9*3, and VKORC1:-1639G>A as the pharmacogenomics features and only some of the non-genetic factors such as age, BSA (only height for some algorithms), and co-morbidities.

Table 2: The Comparison of Model Predictions for IWPC Validation Cohort

Model	Mean±SD of Absolute Error	Median	R^2 (Variance explained)	Intercept	Slope
Sconce	9.50 ± 8.99	6.91	37.7	5.13	0.97
Gage	8.37 ± 7.92	6.47	45.8	2.23	0.93
IWPC	8.39 ± 8.13	6.41	44.5	1.66	0.98
Anderson	8.81 ± 8.11	6.74	41.4	−1.72	1.05

IWPC data set contains 5700 distinct subjects provided by 21 research groups from 9 distinct countries [11], [29]. A study conducted in 2009 [11] on the IWPC data set of PharmGKB [28] states that a pharmacogenetic algorithm including age, race, weight, height, CYP2C9, VKORC1, enzyme inducer status and amiodarone more accurately identifies the patients who require higher doses of 49 mg/week or above and lower doses of 21 mg/week or below compared to a clinical algorithm (including age, race, weight, height, enzyme inducer status and amiodarone) [11].

Gage et al. [22] conducted a study on a derivation cohort of 1015 patients, which proposed another pharmacogenetic model containing 10 significant variables explaining more than half of the variance in the therapeutic warfarin dose (R^2 = 53.1%). The same study also assessed the effect of incorporating genotype information on the prediction of warfarin dosing by excluding the genotype information from the stepwise regression. The dosing algorithm without genetic factors explained only less than half of the variance (R^2=21.5%) [22]. It is worth noting that the accuracy of the model was apparently affected by race: R^2 was calculated 31% in 153 African-American and 57% in 838 Caucasian patients respectively. The equation for warfarin dosing algorithm without genetic factors is as follows:

$$\text{WarfarinDose} = \exp[0.613 + (0.425 * BSA) - (0.0075 * Age) + (0.156 * AfricanAmericanRace) + (0.216 * TargetINR) - (0.257 * Amiodarone) + (0.108 * Smokes) + 0.0784 * DvtPe]$$

(Equation 1) [22]

The non-profit website, http://www.WarfarinDosing.org [80], provides guidance to doctors and other practitioners who initiate warfarin therapy by estimating the therapeutic dose in patients new to warfarin. The web site makes use of the dosing algorithm developed based on the findings of Gage et al. [22].

The FDA altered the warfarin label in 2007 stating that age, BSA, indications for warfarin therapy, concomitant medication, CYP2C9 and VKORC1, were accounted for about 55% of the variance in therapeutic warfarin dose in Caucasian patients [13]. These results were also supported by the IWPC's rather complicated equation which had been provided in the supplementary appendix of Klein et al. [11] . IWPC provides a dosing algorithm in excel format based on this equation [29].

Özer et al. [21] also deduced a linear regression model including CYP2C9*2, CYP2C9*3, VKORC1 −1639, and CYP4F2 (rs2108622) variants, as well as age, which explained 39.3% of the personal differences in the warfarin dose (Table 3).

Table 3: The Parameter Estimates of the Linear Regression Model **[21]**

Variable	Parameter Estimate	Partial R^2	p value
Intercept	8.308		<0.001
VKORC1 −1639A, per A allele	−1.499	0.147	<0.001
CYP2C9 *2 or *3, per variant allele	−1.33	0.19	<0.001
Age (per year)	−0.028	0.028	0.013
CYP4F2, per T allele	0.505	0.028	0.017

2.6 Structural Equation Modelling

Structural equation modeling (SEM) is a statistical approach for estimating and testing causal relationships by using a combination of statistical data and qualitative causal beliefs [14], [15], [17], [19]. The standard SEM (in particular the LISREL model) is basically composed of two models:

1) The measurement model where latent variables (LV) or hidden variables are estimated by using the manifest variables (MV) or observed variables. It shows the relations between latent variables and manifest variables.
2) The structural model where the relations among the latent variables (LV) are assessed. It shows the potential causal dependencies between endogenous (dependent) and exogenous (independent) variables.

The LISREL notation for the full (measurement and structural model) structural equation model is depicted in Figure 4 [17]. As an analogy, confirmatory and exploratory factor analysis models contain only the measurement component, while path diagrams can be regarded as a SEM having only the structural component. SEM enables complex causal relationships to be expressed through linear, non-linear, recursive or non-recursive, hierarchical or non-hierarchical structural equations, to establish a more precise and complete picture of the entire model. There is a growing interest for SEM techniques and in particular they have gained importance in IS research [19]. SEM is used in various multidisciplinary areas which include research, political science, economics, management, marketing, psychology, sociology, and educational research; and the areas are not limited these. Recently, there have been several research papers in Medical domain using SEM techniques [81], [82], [83].

Variable or parameter	*Vector/matrix*	*Typical element*	*Name*	*Explanation*
Variables	$\mathbf{x}$	x_i	x	Vector of observed indicators of ξ
	ξ	ξ_j	ksi	Vector of latent exogenous variables
	$\boldsymbol{\delta}$	δ_i	delta	Vector of errors of measurement of $\mathbf{x}$
	$\mathbf{y}$	y_i	y	Vector of observed indicators of $\boldsymbol{\eta}$
	$\boldsymbol{\eta}$	η_j	eta	Vector of latent endogenous variables
	$\boldsymbol{\varepsilon}$	ε_i	epsilon	Vector of errors of measurement of $\mathbf{y}$
	ζ	ζ_j	zeta	Vector of errors in equations
Parameters	$\boldsymbol{\Lambda}^x, \boldsymbol{\Lambda}^y$	$\lambda^x_{ij}, \lambda^y_{ij}$	Lambda-x, Lambda-y	Matrices of factor loadings for $\mathbf{x}$ and $\mathbf{y}$
	$\boldsymbol{\tau}^x, \boldsymbol{\tau}^y$	τ^x_i, τ^y_i	tau-x, tau-y	Vectors of measurement intercepts
	$\mathbf{B}$	β_{ij}	Beta	Matrix of effects of η_j on η_i
	$\boldsymbol{\Gamma}$	γ_{ij}	Gamma	Matrix of effects of ξ_j on η_i
	$\boldsymbol{\alpha}$	α_j	alpha	Vector of structural equation intercepts
Means of variables	$\boldsymbol{\mu}$	μ_j	mu	Vector of means of $\mathbf{x}$ and $\mathbf{y}$
	$\boldsymbol{\kappa}$	κ_j	kappa	Vector of means of ξ
(Co-)Variances of variables	$\boldsymbol{\Sigma}$	σ_{ij}	Sigma	Covariance matrix of $\mathbf{x}$ and $\mathbf{y}$
	$\boldsymbol{\Phi}$	φ_{ij}	Phi	Covariance matrix of ξ
	$\boldsymbol{\Psi}$	ψ_{ij}	Psi	Covariance matrix of ζ
	$\boldsymbol{\Theta}^{\delta}, \boldsymbol{\Theta}^{\varepsilon}$	$\theta^{\delta}_{ij}, \theta^{\varepsilon}_{ij}$	Theta-delta, Theta-epsilon	Covariance matrices of $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$

Figure 4: The definitions for LISREL notation [17]

SEM is more advantageous to multiple regressions because of several reasons:

a) SEM supports more flexible assumptions (particularly enabling assessment even in the presence of multicollinearity) to be made, SEM uses the confirmatory factor analysis to reduce measurement error by having multiple indicators per latent variable,

b) SEM can handle difficult data (incomplete data, non-normal data, time series with auto-correlated error)

c) SEM is capable of testing models overall rather than factors individually,

d) SEM enable testing of factors across multiple groups,

e) SEM has provision to test models with multiple dependents,

f) SEM is able to model error terms,

g) SEM is more general and more powerful meaning that a variable can act as both independent and dependent variable.

h) Furthermore, SEM is more robust especially in situations where regression is highly exposed to error of interpretation by model misspecification because of its policy of comparing alternative models to assess the relative model fit and selecting the best fitted model among the alternatives.

SEM may be conducted following steps below [18], [84]:

1) Model Specification: The model must be specified correctly based on the type of analysis that is tried to be confirmed. When the correct model is built (i.e. specified), two different types of variables, namely exogenous and endogenous variables are used. The dependence relationship constitutes the difference between these two types of variables. During model specification, a set of theoretically plausible models are suggested so as to assess whether the model proposed is the best among the possible models in the set. The model suggestion is not made only based on the theoretical reasons but the number of data points and the number of parameters must be taken into consideration so that the specified model can be identified. A model is said to be an "identified model" if there is one best value for each parameter that uniquely identifies the model. While specifying the model, two types of relationships can be postulated:
   a) The hypothesized causal relationships (i.e., relationships that do not relay on facts or previous studies) between variables are to be estimated. These relationships are let 'free' to vary.
   b) The known relationships between variables that have been already estimated usually based on previous studies. These are 'fixed' in the model.
2) Estimation of Free Parameters: Parameters are estimated by comparing the actual covariance matrices to the estimated covariance matrices which represent the relationships between variables. This estimation is carried out by using the statistical techniques such as maximum likelihood estimation, weighted least squares or asymptotically distribution-free methods based on maximization of a fit criterion.
3) Assessment of Model Fit: Once the free parameters are estimated, the model is interpreted. The model is rejected or accepted (often being preferred to a competing model) based on statistical tests or assessment of fit measures or indices such as the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI) [85], [86], [87], [88]. These fit measures are calculated by using the resultant matrices expressing the estimated relationships between variables in the model. They essentially evaluate how similar the estimated relationships are to the actual relationships. As there are different measures of fit indices assessing different elements of the fit of the model, a set of fit indices should be used.
4) Model Modification: Although a model is accepted based on assessment of fit indices, it may need to be updated so as to enhance the model fit. There are also modification indices which report the improvement in fit that results from the modification of the model (e.g., adding a new pathway to the model). It is vital that the model modifications are theoretically sensible with regards to the background information.

5) Comparison and Model Selection: The alternate models based on plausible theory are tested and compared to the proposed model and the most appropriate model is selected based on the fit indices such as Browne-Cudeck Criterion (BCC) and the Akaike Information Criterion (AIC) [86].

2.7 Bayesian SEM

Standard practice in implementing SEMs relies on frequentist methods. SEMs are fitted using either generalized least squares procedures or maximum likelihood estimation [14], [18], [89], [90], [91]. To handle more general models and complex data structures, a new approach to factor analysis and structural equation modelling using Bayesian analysis is required [14], [91], [92], [93]. Bayesian approach can be applied to many SEM-type modelling frameworks which include various types of variables such as binary, ordered, categorical, count, and continuous variables expressing diversified interactions such as nonlinearity, multicollinearity among each other. Since the model that is set up for warfarin dosing is also a complex one which involves several types of covariates and missing data, Bayesian SEM is chosen as the most convenient approach for the problem [89], [94]. Other features of the Bayesian approach include linking structural equation concepts to multi-level hierarchical models and estimation of conventional unidentifiable models [93].

2.7.1 Pros and Cons of Bayesian SEM

There are several substantial differences between frequentist and Bayesian approaches:

1) The specification of prior or the model parameters are required by Bayesian SEM. The prior distributions for each of the model unknowns, including the factors from the measurement and structural models and the latent variables are set before the inference. The prior plays an important role on the Bayesian inference because the inference is based on the posterior distributions which rely on both on the likelihood and the prior distribution of the data. Previous studies or theoretical knowledge may provide valuable information about structural relationships and this information can be incorporated into the model through prior distributions. If there is no such information available, vague priors can be made use of. On the other hand, frequentist methods do not specify priors for mean or covariance. Note that, as the sample size increases, the posterior distribution will be influenced less by the prior, and frequentist and Bayesian inferences will become almost the same.

2) The inference method of Bayesian SEM relies on MCMC (Markov Chain Monte Carlo), which draws the instances of the model unknowns (parameters and latent variables) from the corresponding joint posterior distribution by using a computationally exhaustive procedure. As the exact posterior distributions can be inferred for the model unknowns in MCMC, there is no need to rely on large sample assumptions (e.g., asymptotic normality). These exact posteriors can provide a more reasonable measure of model uncertainty in small and medium sized samples but it may last long (e.g., several hours) to obtain enough samples from the posterior.

2.7.2 Markov Chain Monte Carlo and Gibbs Sampling

MCMC methods have been developed to overcome the limitations of Bayesian approaches. These limitations are the incorporation of high dimensional functions, which are used to obtain the posterior distribution, and the augmentation of the observed data with the latent variables or the unobservable data (missing data or the unobservable continuous measurements that actually represent the implicit binary or ordered categorical data).

In MCMC methods, the previous sample values are used to produce the next sample values randomly, generating a Markov chain [14], [20]. The instances of the generated chains are then utilized as samples of the desired distribution after a large number of steps. The quality of the sample gets better as the number of steps increase. When the autocorrelation rate of the generated samples gets higher, MCMC starts to converge slowly. This condition is called slow (or poor) mixing of the MCMC algorithm and causes serious computational problems. Poor mixing can be avoided in certain extent by selecting parameters based on a centered approach. Adaptation of informative priors on the parameters of interest using means taken from an exploratory data analysis with strongly over-dispersed variances also improves convergence [14].

The Gibbs sampling, which is a special implementation of MCMC algorithm uses a slightly different approach and samples from the full conditional posterior distributions. The number of runs (steps) required to reach a stationary chain is a very important factor for the success of MCMC algorithm. Note that, the first 1000 to 5000 samples are discarded (the length of the burn-in period), and then the designated convergence test is utilized to assess whether a stationary chain has been obtained [14], [15].

2.7.3 Software for Bayesian SEM

There are several software applications for the implementation of Bayesian SEM. In this study, two of these applications have been made use of: 1) WinBUGS, and 2) IBM® SPSS® Amos [39]. In the first phase of this thesis, the studies for the proof of concept have been conducted by the freely available WinBUGS software. Since WinBUGS is not fully supported and lacks some graphical and modeling features, the Bayesian modelling and estimation was conducted using AMOS in the second phase.

2.7.3.1 WinBUGS

The WinBUGS[11] is a freely available statistical software package for Bayesian analysis using Markov chain Monte Carlo (MCMC) methods [95]. It succeeded the BUGS (Bayesian inference Using Gibbs Sampling) program which had been launched as a project at the MRC Biostatistical Unit in Cambridge in 1989. The

[11] http://www.mrc-bsu.cam.ac.uk/bugs/

WinBUGS is actually a graphical user interface based on Microsoft Windows version of BUGS program which was jointly developed by a team of UK researchers at the MRC Biostatistics Unit in Cambridge, and the Imperial College School of Medicine at St Mary's of London for Bayesian analysis of complex statistical models using MCMC techniques. The WinBUGS can fit fixed-effect and multilevel models using the Bayesian approach and can handle missing data. Deviance Information Criterion (DIC) can be used for model comparison or goodness-of-fit assessment of the hypothesized model except for the mixture models [96], [97]. Bayesian analysis using the WinBUGS typically involves the following components:

1. Specifying a model that describes the relation between the unknown parameters and the observed data.
2. Specifying prior distributions for the unknown parameters.
3. Obtaining the posterior distributions.
4. Making inference using the posterior distributions.

In the WinBUGS, the model is specified by defining distributions for the data and associated variables. The model specification just involves the calling of the most frequently-used closed-form distributions such as normal, multivariate normal, Bernoulli, binomial, categorical, Poisson etc. which are provided by the WinBUGS. The relationships among variables can be defined by using the <- operator. Prior distributions on the model parameters are also provided in the model specification file, using the ~ operator.

The WinBUGS manual provides more detailed descriptions of the language along with lists of the available logical functions and distributions [96]. The manual also describes how to specify code and run the model and how to evaluate the results. Besides various hypothetical and real life examples which are specified in different type of models (normal hierarchical, multivariate hierarchical, logistic, nonlinear etc.) are provided in volume I and II of WinBUGS examples [98].

The last version of WinBUGS is 1.4.3 which was released in August 2007. The last version is still available but less frequently used. There is an open source successor which is called as OpenBUGS, It essentially contains the core BUGS code with a variety of interfaces and runs under Windows with a very similar graphical user interface to WinBUGS. It can be utilized externally by calling the corresponding programming interface from R language.

2.7.3.2 AMOS

IBM® SPSS® AMOS (Analysis of Moment Structures) is an easy-to-use and powerful software that implements SEM [39], [99], [100]. It provides a graphical interface and visual tools to specify models via path diagrams. Models can be specified, estimated, assessed, and presented in an intuitive diagram based on previous studies and background information. On the other hand, a non-graphical, programmatic method can also be used to specify models. AMOS can also be used to impute missing values and latent scores and utilized for longitudinal studies, multiple-group and reliability analysis. The features of AMOS and how these features have been used will be explained in the upcoming chapters.

CHAPTER 3

DATA AND METHODOLOGY

This chapter introduces the data sets and describes the features of data sets that are used in this study. It also explains the methodology used to achieve the goals of the study.

3.1 Data

Two independent data sets were used for this study: 1) The combined multi-ethnic IWPC data set, and 2) The Turkish data set. The Warfarin Consortium Combined Data Set which consists of 5700 subjects and 68 features are provided by the IWPC were used [11], [29]. The IWPC data set is available at the web site of the Pharmacogenomics Knowledge Base (PharmGKB) [29]. The Turkish data set, on the other hand, contains 107 Turkish patients who were enrolled at the department of Cardiology and Cardiovascular Surgery of Kartal Koşuyolu Education and Research Hospital between April 2009 and January 2010.

3.1.1 IWPC Data

The Warfarin Consortium Combined Data Set (NEJM 2009) which is provided by the International Warfarin Pharmacogenetics Consortium (IWPC) is used [11], [29]. A number of pharmacogenomics research centers have contributed to this data set relating warfarin dosing to a variety of clinical and genetic parameters including genotypes of CYP2C9 and VKORC1. This combined data set consists of 5700 subjects and 68 features. These features are described in detail within the metadata provided in Appendix A. The feature set is preprocessed, reduced and organized as 26 features based on the suggestions of domain experts and the prior knowledge extracted from the literature. Some features are not used in the model. Some new features are derived from original features of the data set. For instance, the comorbidities feature of the original data set is converted into a new feature set of diseases co-occurring in a patient. Some of the features related with medications, and herbal and nutritional intakes are not used at this stage of the study. Medications provided in the original data set, minerals and food supplements, vitamin K will be studied as a future work. The selected features are represented as of 4 data types: continuous, ordinal, categorical and binary. These features are briefly summarized in Table 4. The detailed descriptive statistics (including Mean, Std. Error of Mean, Median, Std. Deviation, Variance, Skewness, Std. Error of Skewness, Kurtosis, Std. Error of Kurtosis) for the selected features are provided in Appendix B.

Table 4: The Selected Features of the IWPC Data Set

No	Feature Name	Label	Description	Type	Missing Rows
1	Age	X1	Represented in bins (starting from 1 corresponding to a decade)	Ordinal	42
2	BSA	X2	The Body Surface Area SQRT(Height*Weight/3600)	Continuous	130
3	Pulmonary Embolism	X4	0 for Absence, 1 for Existence	Binary	0
4	Stroke	X5	0 for Absence, 1 for Existence	Binary	0
5	Deep Venous Thrombosis	X7	0 for Absence, 1 for Existence	Binary	0
6	Cancer	X12	0 for Absence, 1 for Existence	Binary	0
7	Cardiac Failure	X13	0 for Absence, 1 for Existence	Binary	0
8	Hepatic Insufficiency	X16	0 for Absence, 1 for Existence	Binary	0
9	Renal Insufficiency	X17	0 for Absence, 1 for Existence	Binary	0
10	Hypoproteinemia	X18	0 for Absence, 1 for Existence	Binary	0
11	Race	X23	1: White+Caucasian+Hispanic 2: Black or African American 3: Asian (Japanese, Han Chinese, Chinese, Korean, Malay, Indian) 4: Others (Other Mixed Race+Intermediate)	Categorical	0
12	Rifampin	X24	0 for Absence, 1 for Existence	Binary	0
13	Amiodarone	X26	0 for Absence, 1 for Existence	Binary	0

14	Fluconazole	X27	0 for Absence, 1 for Existence	Binary	0
15	Fluvastatin	X29	0 for Absence, 1 for Existence	Binary	0
16	Lovastatin	X32	0 for Absence, 1 for Existence	Binary	0
17	Aspirin	X41	0 for Absence, 1 for Existence	Binary	0
18	Smoking	X42	0 for Absence, 1 for Existence	Binary	0
19	CYP2C9	X46	0 for Absence, 1 for Existence	Binary	145
20	VKORC1:-1639G>A	X48	0 for Absence, 1 for Existence	Binary	1499
21	VKORC1:1173C>T	X49	0 for Absence, 1 for Existence	Binary	2109
22	Therapeutic Dose of Warfarin	Y1	Mean: 30,04 mg/week Sd: 17,37 mg/week	Continuous	172
23	The INR on Therapeutic Dose	Y2	Mean: 2,06 mg/week Sd: 0,9 mg/week	Continuous	732

3.1.2 Turkish Data Set

The data set contains 107 Turkish patients who were enrolled at the department of Cardiology and Cardiovascular Surgery of Kartal Koşuyolu Education and Research Hospital between April 2009 and January 2010 [21]. The patients were selected among 263 individuals who had been using warfarin for at least 4 months and their last three INR measurements were within the therapeutic range (2 to 3) for the stable daily dose. The mean daily warfarin dose was reported as 5.16 ± 1.95 mg (range 1.43–10.00 mg).

22 data features (only relevant anonymous features are considered) which include age, height, weight, gender, smoking, alcohol usage, grapefruit consumption, tea and coffee intake, indication for warfarin, weekly prescribed warfarin dose, hemorrhage/embolisms, other co-morbidities, concomitant medications and pharmacogenomics information were selected. Turkish data set also contains the polymorphisms of CYP4F2 and EPHX1 in addition to CYP2C9*2, CYP2C9*3, VKORC1 -1639G>A and VKORC1 1173C>T that exist in the IWPC data set.

On the other hand, patients who had co-morbidities such as hepatic dysfunction, cancer, advanced heart failure, liver disease and diseases with bleeding tendency or patients who were taking medications that interact with warfarin had been excluded from the original clinical study. Table 5 shows the descriptive statistics for the communized features taken from Turkish data set with respect to the IWPC data set.

Table 5: The Descriptive Statistics for the Communized Features

Continuous Features	**Data Set**			
	IWPC		**Turkish**	
Body Surface Area (BSA) m^2	N (Valid)	4523	N (Valid)	107
	N (Missing)	1177	N (Missing)	0
	Mean	1.8985	Mean	1.8653
	Std. Deviation	0.29896	Std. Deviation	0.17225
Target (International Normalized Ratio) INR	N (Valid)	4870	N/A	
	N (Missing)	830		
	Mean	2.5392		
	Std. Deviation	0.16877		
Therapeutic Dose of Warfarin milligrams/week	N (Valid)	4837	N (Valid)	107
	N (Missing)	863	N (Missing)	0
	Mean	31.4621	Mean	36.1332
	Std. Deviation	16.82898	Std. Deviation	13.66407

Categorical Features	**Data Set**					
	IWPC			**Turkish**		
Age	Values	Frequency	Percent	Values	Frequency	Percent
	10-19 (1)	14	0.2%	10-19 (1)	0	0%
	20-29 (2)	130	2.3%	20-29 (2)	5	4.7%
	30-39 (3)	230	4.0%	30-39 (3)	14	13.1%
	40-49 (4)	540	9.5%	40-49 (4)	19	17.8%
	50-59 (5)	1085	19.0%	50-59 (5)	31	29.0%
	60-69 (6)	1384	24.3%	60-69 (6)	23	21.5%
	70-79 (7)	1570	27.5%	70-79 (7)	15	14.0%
	80-89 (8)	670	11.8%	80-89 (8)	0	0%
	90+ (9)	85	0.6%	90+ (9)	0	0%
	Missing (99)	42	0.7%	Missing (99)	0	0%
Race	Values	Frequency	Percent	N/A		
	White (1)	3239	56.8%			
	African (2)	504	8.8%			
	Asian (3)	1634	28.7%			
	Other (4)	93	1.6%			
	Missing (99)	230	4.0%			
CYP2C9*2	Values	Frequency	Percent	Values	Frequency	Percent
	*1/*1 (1)	4684	82.2%	*1/*1 (1)	86	80.4%
	*1/*2 (2)	737	12.9%	*1/*2 (2)	18	16.8%
	*2/*2 or *2/*3 (3)	125	2.2%	*2/*2 or *2/*3 (3)	3	2.8%
	Missing (99)	150	2.6%	Missing (99)	0	0%
CYP2C9*3	Values	Frequency	Percent	Values	Frequency	Percent
	*1/*1 (1)	4957	87.0%	*1/*1 (1)	87	81.3%

	*1/*3 (2)	498	8.7%	*1/*2 (2)	18	16.8%
	*3/*3 or *2/*3 (3)	91	1.6%	*2/*2 or *2/*3 (3)	2	1.9%
	Missing (99)	150	2.6%	Missing (99)	0	0%
VKORC1 - 1639G>A	Values	Frequency	Percent	Values	Frequency	Percent
	A/A (1)	1485	26.1%	A/A (1)	28	26.2%
	A/G (2)	1470	25.8%	A/G (2)	54	50.5%
	G/G (3)	1246	21.9%	G/G (3)	25	23.4%
	Missing (99)	1499	26.3%	Missing (99)	0	0%
VKORC1 1173C>T	Values	Frequency	Percent	Values	Frequency	Percent
	T/T (1)	1535	26.1%	T/T (1)	27	25.2%
	C/T (2)	1070	18.8%	C/T (2)	55	51.4%
	C/C (3)	986	17.3%	C/C (3)	25	23.4%
	Missing (99)	2109	37.0%	Missing (99)	0	0%
CYP4F2	N/A			Values	Frequency	Percent
				C/C (1)	40	37.4%
				C/T (2)	49	45.8%
				T/T (3)	18	16.8%
				Missing (99)	0	0%
Binary Features	**Data Set**					
	IWPC			**Turkish**		
Pulmonary Embolism	Values	Frequency	Percent	Values	Frequency	Percent
	No PE (0)	5229	91.7%	No PE (0)	101	94.4%
	PE (1)	471	8.3%	PE (1)	6	5.6%
	Missing (99)	0	0%	Missing (99)	0	0%
Stroke	Values	Frequency	Percent	N/A		
	No Stroke (0)	5481	96.2%			
	Stroke (1)	219	3.8%			
	Missing (99)	0	0%			
Cardiac Indications	Values	Frequency	Percent	N/A		
	No CI (0)	3360	58.9%			
	CI (1)	2340	41.1%			
	Missing (99)	0	0%			
Deep Vein Thrombosis	Values	Frequency	Percent	Values	Frequency	Percent
	No DVT (0)	5168	90.7%	No DVT (0)	98	91.6%
	DVT (1)	532	9.3%	DVT (1)	9	8.4%
	Missing (99)	0	0%	Missing (99)	0	0%
Interventions and Surgery	Values	Frequency	Percent	N/A		
	No Surgery (0)	4481	78.6%			
	Surgery (1)	1219	21.4%			
	Missing (99)	0	0%			
Cancer	Values	Frequency	Percent	N/A		

	No Cancer (0)	50	0.9%			
	Cancer (1)	228	4.0%			
	Missing (99)	5422	95.1%			
Cardiac Failures	Values	Frequency	Percent			
	No CF (0)	3553	62.3%			
	CF (1)	879	15.4%			
	Missing (99)	1268	77.8%			
Valve Replacement	Values	Frequency	Percent			
	No VR (0)	3223	56.5%			
	VR (1)	1016	17.8%			
	Missing (99)	1461	74.4%			
Rifampin	Values	Frequency	Percent			
	No Rifampin (0)	2419	42.4%			
	Rifampin (1)	4	0.1%			
	Missing (99)	3277	57.5%			
Amiodarone	Values	Frequency	Percent	Values	Frequency	Percent
	No Amiodarone (0)	3905	68.5%	No Amiodarone (0)	106	99.1%
	Amiodarone (1)	277	4.9%	Amiodarone (1)	1	0.9%
	Missing (99)	1518	26.6%	Missing (99)	0	0%
Fluconazole	Values	Frequency	Percent			
	No Fluconazo. (0)	2409	42.3%			
	Fluconazole (1)	17	0.3%			
	Missing (99)	3274	57.4%			
Fluvastatin	Values	Frequency	Percent			
	No Fluvastatin (0)	2411	42.3%			
	Fluvastatin (1)	12	0.2%			
	Missing (99)	3277	57.5%			
Lovastatin	Values	Frequency	Percent			
	No Lovastatin (0)	2398	42.1%			
	Lovastatin (1)	33	0.6%			

	Missing (99)	3269	57.3%			
Aspirin	Values	Frequency	Percent	N/A		
	No Aspirin (0)	2924	51.3%			
	Aspirin (1)	916	16.1%			
	Missing (99)	1860	32.6%			
Smoking	Values	Frequency	Percent	Values	Frequency	Percent
	No Smoking (0)	2771	48.6%	No Smoking (0)	98	91.6%
	Smoking (1)	449	7.9%	Smoking (1)	9	8.4%
	Missing (99)	2480	43.5%	Missing (99)	0	0%

3.2 Methodology

The proposed framework uses a non-linear approach based on Bayesian Structural Equation Modelling to improve the accuracy and efficacy of warfarin dosing while using the proper subset of genetic and non-genetic features provided in IWPC and Turkish data sets [15], [18], [29]. An iterative method is applied until the best fitted model is obtained which involved several steps of data pre-processing, data analysis, model construction, Bayesian inference, testing and validation (Figure 5).

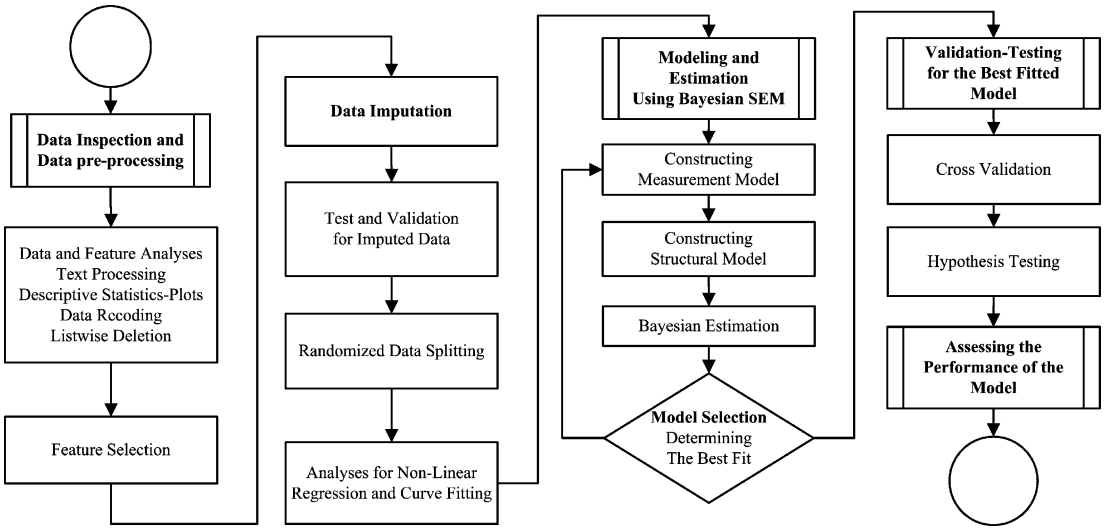

Figure 5: Steps followed to establish the modelling and estimation framework for warfarin dosing

Initial steps include some extensive data processing and analysis tasks such as feature selection, missing value analysis and data imputation. Both the IWPC and Turkish data sets were obtained in excel format. Therefore, Excel and IBM SPSS Statistics 21 [101], [102] were used extensively for the initial data processing steps such as data inspection and data re-coding. The preliminary results indicate that some of the attributes seem to have less impact (such as X6: cardiac indications).

Therefore, it is better to select the more relevant attributes in the first place and then apply the Bayesian SEM on a reduced number of attributes [14], [100], [101]. The less relevant attributes could then be incorporated into the model in later steps. This approach was suggested to improve the efficacy and decrease the complexity.

Once the data set is ready for modelling and inference, Bayesian SEM is applied. In this chapter, the features of the data sets used are expressed and the data pre-processing methods are elaborated in the first place. In the second place, the actions taken for conducting Bayesian SEM and estimation are explained.

CHAPTER 4

PROPOSED FRAMEWORK FOR WARFARIN DOSING

This chapter elaborates the steps taken to establish the proposed Bayesian modeling and estimation framework for pharmacogenomics driven warfarin dosing based on the on the methodology provided in the previous chapter. The major steps are as follows:

a) Data Inspection and data pre-processing,
b) Missing value analysis and data imputation,
c) Feature selection,
d) Bayesian modelling and estimation using AMOS,
e) Assessing the performance of the model.

4.1 Data Inspection and Data Pre-processing

The first step was to inspect the data set and analyze its features. Several descriptive statistics were obtained to assess the nature of the features. There were basically 3 types of features: continuous, binary and categorical. Apart from continuous binary and categorical features, there were also text based features for concomitant diseases that had been typed in free format giving a list of diseases co-occurring in a patient separated by commas. These features were processed by some scripts and converted into a set of binary features such as cancer, diabetes, cardiac failure, hepatic failure, renal insufficiency, and hypo-proteinemia based on the suggestions of the domain experts and frequency. Similar processing was performed for the features related with indications. The features related with medications, herbal and nutritional intakes were also processed in a similar way, but these features could not be made much use of due to the lack of data.

Consequently, an initial set of 68 features were obtained for IWPC data set and all features were inspected using the descriptive statistics according to the type of the feature [101]. Continuous features were assessed with mean, standard deviation, skewness and kurtosis to understand the characteristics of the distribution. Categorical and binary features were analyzed based on the frequencies. Categorical features having n possible values were recoded to m binary features where only one of them was active for the corresponding data row. For this reason, the categorical features such as Age and Race were initially assessed with histogram plots. For

instance, the histogram of the features Age is provided in Figure 6.

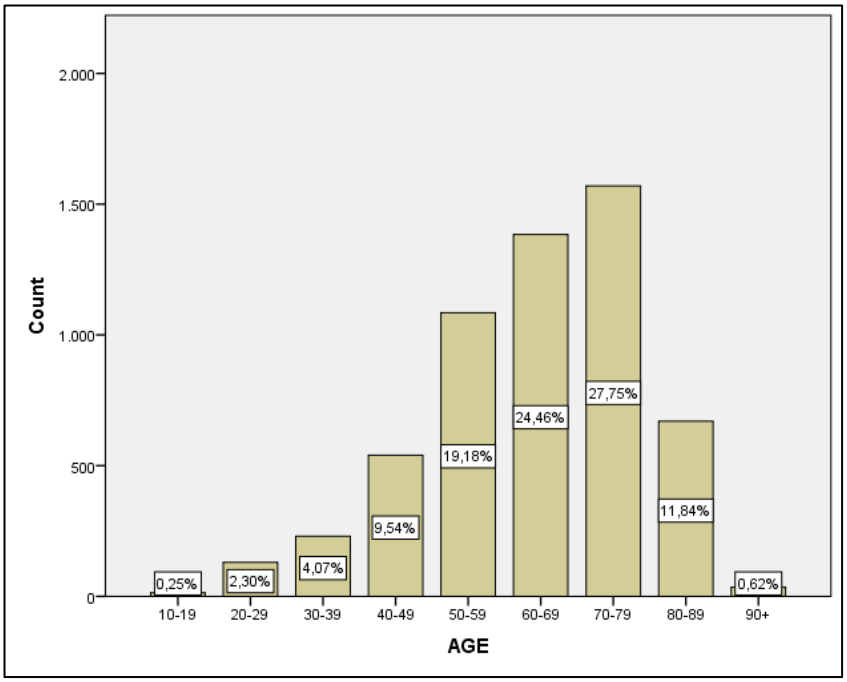

Figure 6: The Histogram Plot for Age

On the other hand, the continuous features were assessed against the target variable (Therapeutic Dose of Warfarin) to reveal their distribution characteristics. Such a plot is provided for the variable INR on Therapeutic Dose against the target variable Therapeutic Dose of Warfarin. The plot (Figure 7) exhibited a non-linear relationship which roughly indicated that INR on Therapeutic Dose could be accounted for the variation in warfarin dose. Note that, the continuous features were normalized to eliminate kurtosis using natural logarithm. For instance, the therapeutic dose of warfarin (Y1) was also normalized using the log function.

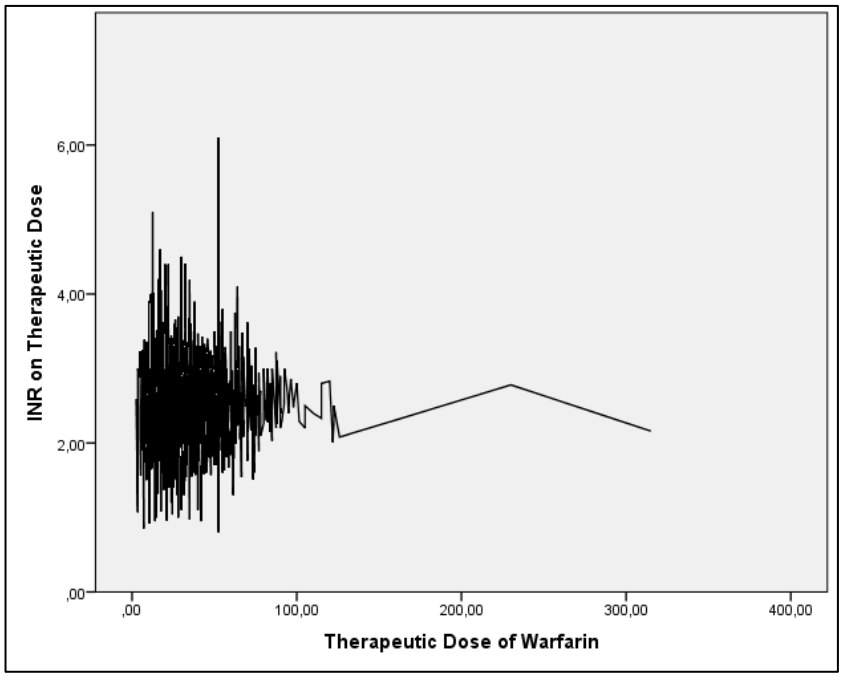

Figure 7: INR on Therapeutic Dose vs Therapeutic Dose of Warfarin

Although IWPC data set originally contains 5700 subjects, the subjects who were marked as patients that reached a stable dose of warfarin (provided as a feature in the data set) and had an INR of 2 to 3 are selected (4523 rows) at the very first step of data processing within the scope of this study. Some genetic and non-genetic features such as VKORC1 1542, VKORC1 4451, and taking azoles were not utilized in the model because of the high percentage of missing rows. Some new features were derived from original features of the data set. For instance, the co-morbidities feature of the original data set which had been expressed as a list of diseases, were converted into a set of new binary features that correspond to co-occurring diseases (cancer, diabetes, cardiac failures, valve replacement). Similar processing was performed for the features related with indications. Some of the features related with medications, and herbal and nutritional intakes were left for future studies.

Since the non-synonymous SNPs in CYP2C9 (*2 and *3) are both aligned with reduced enzyme activity, they were treated as separate categorical factors. The genotypes of CYP2C9*2 (*1/*1, *1/*2, *2/*2) were coded as a categorical factor having values 1 (wild), 2 (heterozygote) and 3 (mutant) respectively. Similarly, the genotypes of CYP2C9*3 (*1/*1, *1/*3, *3/*3) were coded as 1 (wild), 2 (heterozygote) and 3 (mutant). The individuals having both of the CYP2C9*2 and CYP2C9*3 mutations (*2/*3) were accepted as mutant.

The genotypes of VKORC1 -1639G>A (A/A, A/G, G/G) were coded as a categorical factor having values 1 (mutant), 2 (heterozygote) and 3 (wild). Similarly, VKORC1 1173C>T (T/T, T/C, C/C) were coded as a categorical factor having values 1 (mutant), 2 (heterozygote) and 3 (wild).

The CYP2C9 (X46) stands for the categories of variants of CYP2C9. Recall that, the variants of CYP2C9 have been shown to influence warfarin dose as well as affecting the clearance of several other drugs [22], [33], [70]. Combined QC CYP2C9 attribute

from the original data set which contained the combined separate genotypes for *2 and *3 is used. CYP2C9*1 metabolizes warfarin normally, CYP2C9*2 causes a reduction in warfarin metabolism by 30%, while CYP2C9*3 causes a reduction in warfarin metabolism by 90%. Because warfarin given to patients with *2 or *3 variants will be metabolized less efficiently, the drug will remain in the circulation longer, so lower warfarin doses will be needed to achieve anticoagulation. The prevalence of each variant varies by race; 10% and 6% of Caucasians carry the *2 and *3 variants, respectively, but both variants are rare (< 2%) in those of African or Asian descent. These separate genotypes were initially represented as single CYP2C9 diplotypes with possible values *1/*1 (1), *1/*2 (2), *1/*3 (3), *2/*2 (4), *2/*3 (5), *3/*3 (6). Therefore, there were 6 categories having frequencies 4164, 737, 498, 56, 69 and 22 out of 5700 respectively. The category (1) represented the situation without polymorphism and had the highest frequency as expected. Note that, diplotype information was not reported for 145 subjects.

The VKORC1:-1639G>A (X48) represents one of the most well-known SNPs of VKORC1 [26], [67]. It is a polymorphism in the promoter region of VKORC1 that is believed to be the causative SNP for the low dose phenotype. This polymorphism alters the VKORC1 transcription factor binding site and the activity of the G allele is increased by 44% over the activity of the A allele [67]. The changes in gene expression presumably lead to fewer functional copies of the mature VKORC1 protein, which is the rate limiting enzyme in the vitamin K cycle. This polymorphism has pronounced differences in its frequency by ethnic group as it is actually the majority allele (around 90%) in Asian populations and appears to explain the lower warfarin dose requirement for individuals of Asian descent. This variant is also quite common in Caucasians, with an allele frequency typically around 40%.

The attribute can take the values A/A (1), A/G (2), G/G (3) where G/G (3) actually represents the condition without polymorphism.
The categorical frequencies for X48 are 1485, 1470, and 1246 respectively. For 1489 subjects, a category is not stated.

The VKORC1:1173C>T [13] (X49) is the attribute that represents the SNP in the first intron of VKORC1, and is in near perfect linkage disequilibrium with VKORC1:-1639G>A (X48). This SNP is highly observed in Asian populations (Japanese, Han Chinese) can take the values T/T (1), C/T (2), C/C (3) where T/T (1) actually represents the condition without polymorphism. In the original data set, the frequencies of X49 with respect to the corresponding categories are calculated as 1535, 1070, and 936 respectively. There are 2109 unreported values for this attribute.

Therapeutic Dose of Warfarin (Y1) is the target manifest variable. It is a continuous variable with a mean of 30.0652 and standard deviation of 17.34. There are 172 missing values.

The International Normalized Ration (INR) on Therapeutic Dose (Y2) is the second target manifest variable. It is also a continuous variable with a mean of 2.1213 and standard deviation of 0.96. Note that, the whole data set is not suitable for the implementation of Bayesian SEM. Therefore, a series of data pre-processing and data

imputation techniques are applied. Consequently, an imputed data of 4523 rows and 24 columns are obtained.

One third of the imputed data is put aside for testing purposes using cross validation. The remaining two thirds are used for Bayesian estimation.

4.2 Missing Value Analysis and Data Imputation

Proper handling of missing values is critical for the soundness of any analyses [103], [104], [105], [106], [107]. Missing data reduce the representativeness of the sample and can therefore distort inferences about the population. Various techniques can be used to handle missing data. One trivial way to deal with the missing values is to simply remove the rows having missing values from the working data set (listwise or pairwise deletion). If this approach were applied for IWPC dataset, the number of rows would be reduced to 108 from 5700.

Other technique which is mainly used is imputation (filling in or rectangularization). Imputation is the substitution of a value for a missing data point or a missing component of a data point. Once all missing values have been imputed, the dataset can then be analyzed using standard techniques for complete data. However, it is not trivial to impute missing data, if there are more than one missing covariates for a sample. Many features of IWPC data set contained missing values. For instance, the features X1:Age and X2:BSA had relatively small number of missing values which were 42 and 130 respectively but the features for indications, co-morbidities and -medications contained significantly large number of missing values. The summary of missing value analyses and corresponding frequencies (Table 6) demonstrated that the percent missing ranges from 0.4% to 94% with respect to the factor.

Different approaches were followed according to the percent of data missing [103], [104]. Therefore, 17 rows of race and 33 rows of age were listwise deleted and eventually 4473 rows are left for further data analyses and modeling. For factors having more than 5% and less than 50% missing, multiple imputation was used accordingly. Factors (cancer 94,0%, fluvastatin 53,1%, rifampin 53,1%, fluconazole 53,0%, lovastatin 52,9%) having more than 50% missing values were totally excluded from the study since multiple imputation performs poorly in these cases.

Table 6: Statistics for Missing Values

	Missing		Valid N	Mean	Std. Deviation
	N	Percent			
CANCER	4251	94.0%	272		
FLUVASTATIN	2400	53.1%	2123		
RIFAMPIN	2400	53.1%	2123		
FLUCONAZOLE	2397	53.0%	2126		
LOVASTATIN	2392	52.9%	2131		
VKORC1 1173C>T	1917	42.4%	2606		
SMOKING	1686	37.3%	2837		
TDoW-wrt-INRonTD	1375	30.4%	3148	1.4786	.21149
DIABET	1375	30.4%	3148		
AMIODARONE	1111	24.6%	3412		
ASPIRIN	1048	23.2%	3475		
VKORC1 -1639G>A	824	18.2%	3699		
VALVE REPLACEMENT	804	17.8%	3719		
TARGET_INR	754	16.7%	3769	2.5525	.18264
CARDIAC FAILURES	590	13.0%	3933		
AGE	33	.7%	4490		
RACE	17	.4%	4506		
CYP2C9*3	0	0.0%	4523		
CYP2C9*2	0	0.0%	4523		
INTERVENTIONS AND SURGERY	0	0.0%	4523		
DVT	0	0.0%	4523		
CARDIAC INDICATIONS	0	0.0%	4523		
STROKE	0	0.0%	4523		
PE	0	0.0%	4523		
BSA	0	0.0%	4523	1.8985	.29896

In the first stage of the study, the multiple imputation was implemented using Amelia toolbox [104]. Amelia II's Expectation Maximization Bootstrap (EMB) algorithm allows users to impute incomplete data sets. Thus, the analyses which require complete observations can appropriately use all the information present in a dataset with missingness, and avoid the biases, inefficiencies, and incorrect uncertainty estimates that can result from dropping all partially observed observations from the analysis.

The result of imputation was validated by comparing the distribution density of the original data set against the imputed data set. The suggestions of the domain experts and the prior knowledge gathered from the literature were also used to validate the imputed values. Based on the prior knowledge that the polymorphism VKORC1:1173C>T is highly observed in Asian population, the imputed values that

correspond to Asian individuals were checked whether they were polymorphic [35]. The rows that contain over-imputed values are treated as outliers and deleted from the imputed data set. An imputed data of 4523 rows and 24 columns are obtained consequently.

In the second stage of the study, the multiple imputation techniques of SPSS Statistics were utilized. SPSS Statistics makes use of a Fully Conditional Specification (FCS) method based on MCMC algorithm for multiple imputation [101], [103]. The missing patterns of the data set are analyzed. The features such as cancer having a high percentage of missing values were excluded from the study. For the rest of the independent features which were binary variables essentially, multiple imputation was performed using logistic regression on the recoded data set. The number of imputations was set to 5. The number of between-imputation iterations was assigned to 200 (a data set was saved every 200^{th} iteration). In order to assess the accuracy of the imputations, some of the original data are randomly deleted, multiple imputation was carried out and the imputed values were compared to the actual values. The comparison was made for randomly selected 100 cases and the proportion of the imputed categories that matched the true categories was calculated. The comparison results indicated that the imputations were ≈96% accurate.

4.3 Feature Selection

Factors were excluded from the analysis for several reasons other than being missing. For instance, race was discarded due to two main reasons: a) Multicollinearity: Race was shown to be strongly correlated with the genetic factors in IWPC data set. b) Communization: Race was obviously not a factor in Turkish data set and two data sets were needed to be communized for comparison purposes. Note that, multicollinearity does not reduce the reliability of the model or the overall predictive power but can affect the results of the individual predictors. VKORC 1173C>T was also extracted from the model because of multicollinearity. The result of the linear co-variance analysis between VKORC1: -1639G>A and VKORC1: -1639C>T where the therapeutic dose of warfarin (Y) was the dependent variable indicated that two genetic factors were highly correlated which is also consistent with the fact that two mutations are in near perfect linkage disequilibrium with each other (Table 7).

Table 7: Linkage Disequilibrium for VKORC1 SNPs

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	23.538[a]	6	3.923	109.953	.000
Intercept	84.505	1	84.505	2368.485	.000
x48	.219	2	.110	3.071	.047
x49	.719	2	.359	10.072	.000
x48 * x49	.222	2	.111	3.115	.045
Error	49.701	1393	.036		
Total	3067.226	1400			
Corrected Total	73.239	1399			

a. R Squared = .321 (Adjusted R Squared = .318)

In data mining, pattern recognition, machine learning and statistics, feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique of selecting a subset of relevant features for building robust learning models. The removal of the most irrelevant and redundant features from the data set helps to improve the performance of learning models. The features can be selected to be mutually far away from each other (minimum redundancy), while they still have "high" correlation to the classification variable (maximum relevance). One of the algorithms in this category is so-called **minimum-redundancy-maximum-relevance selection** (mRMR) [108], [109] .

In this study, mRMR based on Mutual Information Difference (MID) criterion is used to obtain the most relevant feature set with minimum classification error. For mutual information based feature selection methods, the algorithm is asserted as it leads to better results with categorical data compared to the continuous data. Our data set contains mostly categorical or binary data. A binning algorithm is utilized before the application of mRMR for the features such as X1: BSA, X10: Target INR etc.

Note that, the feature selection algorithm has been run separately for the two target variables Y1: Therapeutic Dose of Warfarin and Y2: INR on Therapeutic Dose. The feature selection algorithm is applied heuristically in groups against Y1 and Y2. For instance, the medications and intakes data (x24-x45) are processed separately in conjunction with Y1 and Y2. The results are provided below in order:

1. The selected features related with Medications and Intakes are listed in descending order according to their importance as follows:

- X42-Smoking
- X41-Aspirin
- X24-Rifampin
- X26-Amiodarone

- X27-Fluconazole
- X29-Fluvastatin
- X32-Lovastatin

2. The selected features related with physical features, co-morbidities and indications are listed in descending order according to their importance as follows:

- X2-BSA
- X1-AGE
- X23-RACE
- X14-CARDIAC FAILURE
- X10-TARGET-INR
- X7-DVT
- X4-PE
- X5-STROKE
- X12-CANCER
- X16-HEPATIC FAILURE
- X17-RENAL INSUFFICIENCY
- X18-HYPOPROTEINEMIA

3. The pharmacogenetic features are listed in descending order as follows:

- X46-CYP2C9
- X48-VKORC1 -1639
- X49-VKORC1 1173

68 attributes were reduced to 24 attributes (including the target variables Y1 and Y2) as a result of the feature selection algorithm.

The feature selection was also applied by using SPSS Statistics on the processed and imputed IWPC data set and the similar set of features was obtained (Table 8): Age, BSA, PE/DVT, Amiodarone, Smoking, CYP2C9*2, CYP2C9*3, VKORC1 -1639G>A, Target-INR and Warfarin Dose. The data used in modelling consisted of 4473 rows and 10 features. 1491 subjects were randomly selected to form the validation cohort. The remaining 2982 rows were used as the training cohort for the IWPC data set.

Table 8: The Results of Principal Component Analyses by using SPSS Statistics

	Component								
	1	2	3	4	5	6	7	8	9
AGE	-.519	-.066	.435	-.236	.213	-.134	.165	.196	.231
BSA	.128	-.166	-.414	.248	-.272	.207	.568	-.023	-.014
RACE	.692	-.075	.123	.113	-.157	-.158	-.150	-.307	.090
AMIODARONE	-.450	-.095	-.399	.216	.169	.084	-.317	.367	-.035
CYP2C9*2	-.231	.380	.177	-.394	.285	.227	.205	-.154	-.032
CYP2C9*3	-.188	-.026	.037	-.440	-.091	.528	.109	-.096	-.020
VKORC1 -1639G>A	.752	.121	.227	.108	.468	.166	-.159	.141	.040
VKORC1 1173C>T	.754	.113	.213	.128	.459	.168	-.153	.154	.049
Therapeutic Dose of Warfarin	.606	-.030	-.285	.196	.119	.135	.419	.090	.002
INR or Therapeutic Dose	.021	.255	.015	.016	-.043	-.466	.234	.601	-.248
INTERVENTIONS AND SURGERY	-.127	.892	-.055	.127	-.155	.011	-.097	-.125	.126
CARDIAC INDICATIONS	-.660	-.181	-.052	.074	.469	-.094	.104	-.001	.165
CARDIAC FAILURES	-.364	-.015	-.523	.380	.226	.021	-.222	-.149	-.021
VALVE REPLACEMENT	-.108	.901	.015	.100	-.107	.032	-.018	-.071	.107
STROKE	-.022	-.168	.447	.544	-.250	.192	.005	-.002	-.368
DVT	.436	-.095	-.269	-.462	-.249	.137	.094	.311	.286
PE	.396	-.001	-.200	-.412	.050	-.375	-.164	-.036	-.201
TARGET INR	-.016	.627	.013	.026	-.066	-.022	.180	.130	-.314
ASPIRIN	-.225	-.110	.442	.205	-.170	.380	-.016	.315	.071
SMOKING	.099	.227	-.209	.220	-.100	.136	-.111	.219	.540
LOVASTATIN	-.037	-.134	.201	-.031	-.569	-.175	-.317	.102	.215
DIABETES	.028	-.058	.271	.322	.093	-.417	.438	-.189	.394

Note that, Turkish data set has no missing values for the features that are included in this study and the similar features used for IWPC are also utilized for the Turkish data set: Age, BSA, Indications, CYP2C9*2, CYP2C9*3, CYP4F2, VKORC1 -1639G>A, VKORC1 1173C>T, Warfarin Dose. The Turkish data set is randomly split up into two cohorts as well: Training cohort (72 patients) and validation cohort (35 patients).

4.4 Bayesian Modelling and Estimation with AMOS

Structural equation modeling (SEM) was used as the modelling and estimation framework since it covers a broad range of approaches from linear regression to Bayesian inference including the confirmatory factor analysis [14], [20]. Bayesian approach which allows nonlinearity, missing data, various types of observed variables (mixed categorical, binary, and continuous) was preferred to other approaches due to its flexibility and strength in estimation and analysis for SEM frameworks. Bayesian SEM, therefore, was considered to be the most convenient approach for the complex warfarin dosing problem containing several types of covariates and missing data.

The initial hypothetical structural equation model for warfarin dosing initially consisted of 24 observed variables –two of them are dependent observed variables Y1 and Y2- was expressed in LISREL (LInear Structural RELations) notation (Figure 8). Note that, the model was constructed intuitively based on the previous

studies and suggestions of the domain experts. In LISREL notation, the observed (manifest) variables are shown in rectangles whereas the latent variables are shown in ellipses or circles [19], [20]. Remember that a latent variable is a variable that cannot be observed directly and must be inferred from measured variables. The measurement model defines the constructs (latent variables) that the model will use, and assigns observed variables to each latent variable. The arrows between the latent variables indicate these structural connections. Latent variables are also implied by the covariance values among two or more measured variables; and called as factors (i.e., factor analysis), constructs or unobserved variables. Therefore, the latent variables are important for the factor analysis. Error terms ("disturbances" for latent variables) are also included in the SEM diagram, represented by "e".

In the hypothetical model (Figure 8), Physical Features (ξ1), Indications (ξ2), Co-Morbidities (ξ3), Inducers (ξ4), Inhibitors (ξ5), Other Intakes and Medications (ξ6), CYP2C9 (ξ7), VKORC1 (ξ8), Ethnicity (ξ9) are the latent exogenous variables and are represented in circles. Warfarin Dose (η1) is the latent endogenous variable and is also represented in circle. Exogenous observed variables are represented in rectangles and with the letter X. Age (X1) and BSA (X2) were assigned to the latent variable, Physical Features (ξ1). PE (X4), Stroke (X5), Cardiac Indications (X6), DVT (X7), and Interventions and Surgery (X8) were set to Indications (ξ2). Cancer (X12), Diabetes (X13), Cardiac Failures (X14), and Valve Replacement (X15) were represented as factors of Co-Morbidities (ξ3). Rifampin (X25) was linked to the latent variable, Inducers (ξ4). Amiodarone (X26), Fluconazole (X27), Fluvastatin (X29), and Lovastatin (X32) were assigned to Inhibitors (ξ5). Aspirin (X41) and Smoking (X42) were set to Other Intakes and Medications (ξ6). CYP2C9*2 and CYP2C9*3 were assigned to CYP2C9 (ξ7). VKORC1 -1639G>A (X48) and VKORC1 1173C>T (X49) were represented as the factors of VKORC1 (ξ8). Finally, Race (X23) was linked to Ethnicity (ξ9). Target-INR (X10) was directly linked to the latent endogenous variable, Warfarin Dose (η1). Observed variables were handled according to the type of the variable. Age (X1), for instance, is an ordered categorical variable, which was treated as observations that were coming from a hidden continuous normal distribution with a threshold specification. The covariance values between ξ1-ξ2, ξ2-ξ3, ξ4-ξ7, ξ5-ξ7, ξ7-ξ8, ξ8-ξ9 and ξ7-ξ9 are investigated. The covariance between two variables equals to the correlation times the product of the variables' standard deviations. The covariance of a variable with itself is the variable's variance. By that, the hidden correlations between latent factors (ξ7: CYP2C9 and ξ8:VKORC1) were studied.

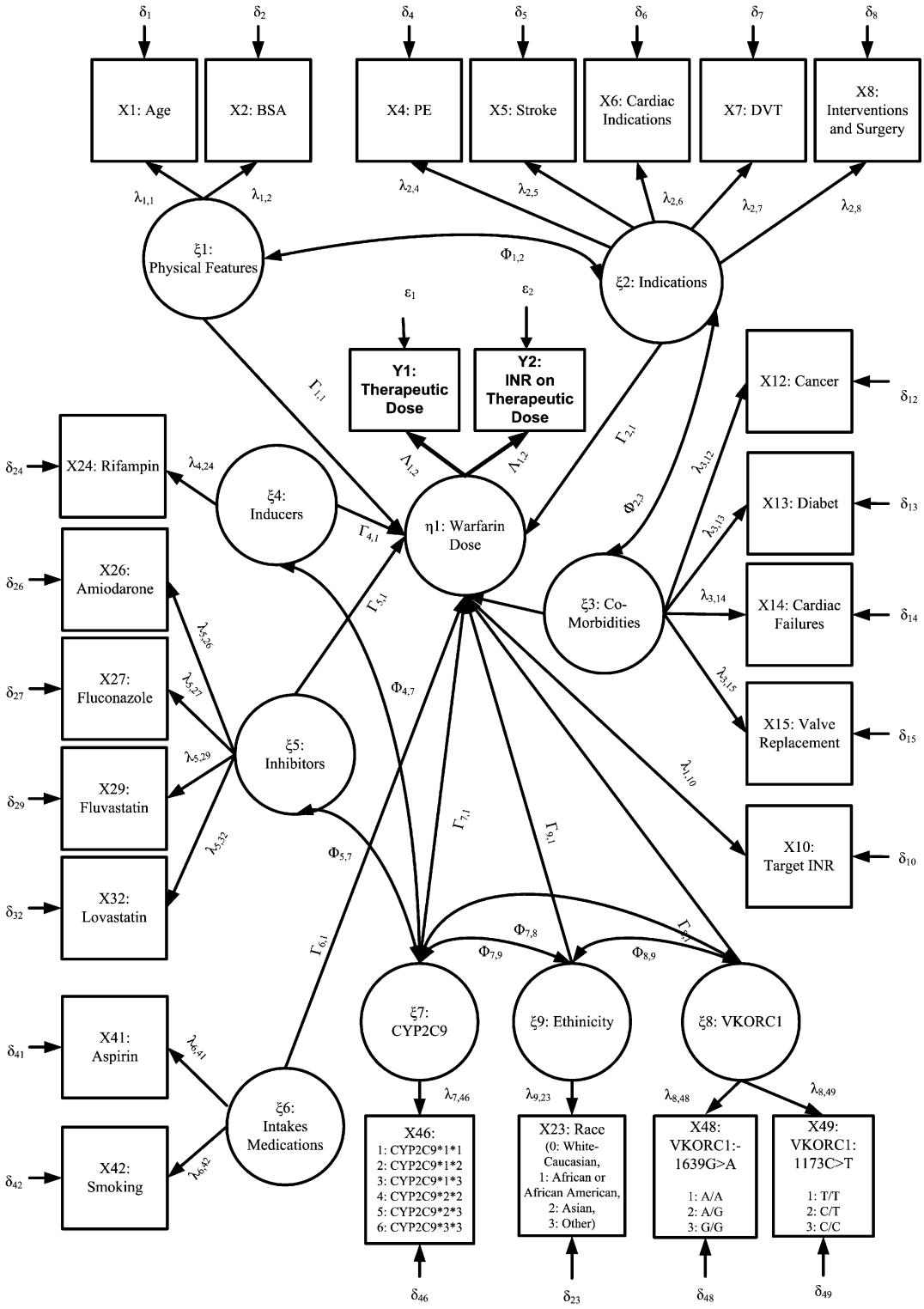

Figure 8: Hypothetical SEM for Warfarin Dosing

However, in practice the models were constructed according to the factors existed in the corresponding data set in an incremental fashion. For instance, while the model for Turkish data set included CYP4F2, the model for IWPC data set did not take CYP4F2 into account. Therefore, the final (recognized as the best fitted model according to the goodness-of-fit tests for a set of factors) model was named after the

subset of factors incorporated from the corresponding IWPC or Turkish data sets (Table 9). Note that, this task was performed separately for each of the models. Model-1 could eventually be identified with 10 factors from IWPC data set. Similarly, Model-2 included 8 factors from Turkish data set. Model-3 and Model-4 were established as the models that contain the same factors from the corresponding data sets.

Table 9: Data Subsets and Corresponding Model Names wrt. the Factors Included

Description	**Factors Included**	**Name of the Bayesian Model based on the Subset**
10 factors from IWPC Data Set	AGE, BSA, PE/DVT, AMIODARONE, TARGET-INR, SMOKING, CYP2C9*2, CYP2C9*3, VKORC1 -1639G>A, Warfarin Dose	Model-1
8 factors from Turkish Data Set	AGE, BSA, Indications (DVT/PE/Cardiac Indications), CYP2C9*2, CYP2C9*3, VKORC1 -1639G>A, CYP4F2, Warfarin Dose	Model-2
5 communized factors from IWPC Data Set	AGE, CYP2C9*2, CYP2C9*3, VKORC1 -1639G>A, Warfarin Dose	Model-3
5 communized factors from Turkish Data Set	AGE, CYP2C9*2, CYP2C9*3, VKORC1 -1639G>A, Warfarin Dose	Model-4

Bayesian estimation and analysis was applied separately for each of the models utilizing the IBM® SPSS® AMOS™ (Analysis of Moment Structures) v21 software [39]. The graphical representation of the final model based on 10 communized factors (Model-1) that was identified and converged in AMOS is provided in Figure 9. Bayesian SEM analysis and inference in AMOS which is based on MCMC involved the following several steps:

a) Model specification and identification,
b) Selecting prior distributions and the admissibility test,
c) Defining analysis properties (burn-in, bootstrap),
d) Estimating explicit means and intercepts,
e) Testing for convergence,
f) Model selection based on model fit.

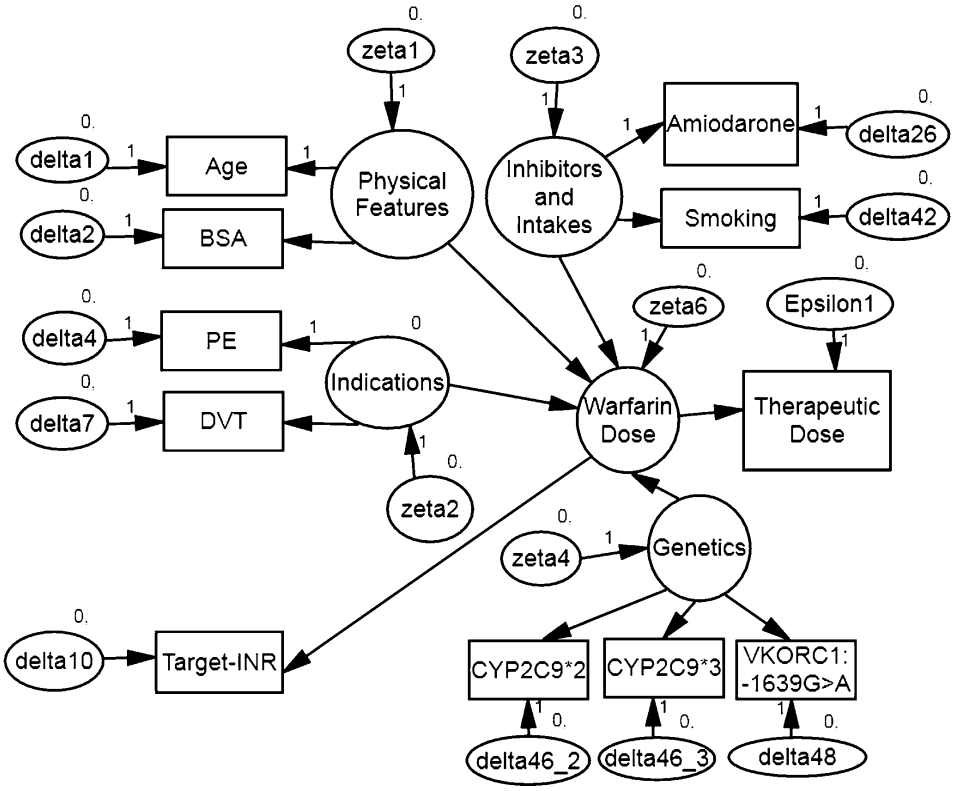

Figure 9: AMOS Representation for Model-1

4.4.1 Model Specification and Identification

Once the data analysis and data processing steps were accomplished, Bayesian Modelling and Estimation step was carried out. AMOS provides utilities for reading the data from various resources. The data can easily be obtained from the excel file or SPSS Statistics data file (.sav) [39], [101]. In this study, the data files in Excel format were converted to SPSS data files to facilitate further data analysis by using statistical tools in SPSS.

The hypothetical model was easily drawn using the graphical tools provided at Diagram menu. Observed and latent variables were sketched according to the corresponding figures (circle for latent, rectangle for observed) by just drag and drop movements. The objects drawn, then, were named according to the corresponding labels of the variables in the data set. The object properties were easily assigned by right-clicking the selected variable and setting the values in 'Object Properties' window opened from the pop-up menu. An ellipse was drawn to represent the error variable for each observed variable. Single-headed arrows that point from the exogenous, or predictor, variables to the endogenous, or response, variables were drawn to complete the model. Note that, endogenous variables should have at least one single-headed path pointing toward them. On the other hand, exogenous variables have only single-headed paths in outward direction but do not receive any arrows.

After the model was created and specified, the identification process took place. If any of the regression weights were impossible to be estimated, then the model would be un-identified. The identification problems were corrected by fixing either the regression weight applied to error variable or the variance of the error variable itself for the corresponding model variable, Regression weight was fixed at 1 for some of the variables which yielded to the same estimates as linear regression.

4.4.2 Application of Standard SEM

At the very initial phase of the study, the standard SEM was exercised to model the warfarin dosing algorithm. This exercise was carried out to gain a basic understanding of SEM and to assess the basic functionality and validity of SEM approach. It was also advised to perform a maximum likelihood analysis for comparison purposes before performing the Bayesian SEM [39]. The linear equation used to formulate the linear measurement model is given in Table 10. SEM package in R was utilized for the exercise. The coefficients obtained by 2SLS estimation were found to be consistent with the values given in the reference study since the standard SEM is basically an extension of the general linear model (GLM) and yielded to similar results with MLE as shown in Table 11. AMOS enables to perform MLE or GLM analysis by selecting the corresponding option in Analyses Properties window.

Table 10: The Measurement Equation for the Standard SEM

$\mathbf{Y=\alpha_1X_1+\alpha_2X_2+\alpha_3X_3+\alpha_4X_4+\alpha_5X_5+\alpha_6X_6+\alpha_7X_7+\alpha_8X_8+\alpha_9X_9+\alpha_{10}X_{10}}$
Y= Warfarin dose,
X_1=Age, X_2=BSA (Body Surface Area), X_3=Race ,X_4=VKORC1 genotype,
X_5=CYP2C9 genotype, X_6=Target INR, X_7=Taking Enzyme Inducer, X_8=Taking
Amiadorone, X_9=Smoking, X_{10}=DVT/PE

Table 11: The Results of Standard SEM for IWPC Data Set

	Uc[a]		SC[b]	t	Sig.	95.0% Confidence Interval for B		Correlations		
	B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part
(Constant)[c]	1.519	.029		53.116	0.000	1.462	1.575			
VKORC1 -1639G>A-MUTANT	-.237	.010	-.451	-23.997	.000	-.256	-.218	-.355	-,485	-,407
AGE	-.034	.002	-.264	-15.103	.000	-.039	-.030	-.281	-,330	-,256
VKORC1 -1639G>A-HETERO	-.113	.008	-.276	-14.972	.000	-.128	-.098	-.135	-,327	-,254
BSA	.154	.012	.222	12.778	.000	.130	.177	.315	,283	,217
CYP2C93_3	-.240	.034	-.144	-6.975	.000	-.307	-.172	-.172	-,159	-,118
CYP2C93_2	-.131	.012	-.188	-10.923	.000	-.155	-.108	-.189	-,245	-,185
AMIODARONE	-.104	.014	-.129	-7.581	.000	-.131	-.077	-.138	-,173	-,128
CYP2C92_2	-.070	.009	-.132	-7.635	.000	-.089	-.052	-.075	-,174	-,129
CYP2C92_3	-.124	.027	-.096	-4.640	.000	-.176	-.071	-.159	-,107	-,079
SMOKING	.037	.010	.062	3.600	.000	.017	.057	.138	,083	,061
DVT/PE	.037	.010	.062	3.581	.000	.017	.058	.158	,083	,061

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
.681	.463	.460	.14763	.004	12.824	1	1869	.000

The results of MLE applied on the curated IWPC data set for 10 factors (Subset-3). The dependent Variable Y is Warfarin Dose. X4-PE was excluded during stepwise regression.
a.Unstandardized Coefficients
b.Standardized Coefficients
c.Intercept

Similarly, the Standard SEM was also applied for the Turkish data set containing 5 factors (Model-4) and consequently the consistent results with the referenced study [21] were obtained as tabulated in Table 12.

Table 12: The Results of Standard SEM for Turkish Data Set

	UC[a]		SC[b]	T	Sig.	95.0% Confidence Interval for B		Correlations		
	B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part
(Constant)	4.113	.131		31.419	.000	3.853	4.373			
VKORC1 - 1639G>A-MUTANT	-.676	.096	-.683	-7.043	.000	-.867	-.486	-.409	-.578	-.527
CYP2C9*3_3	-1.001	.235	-.323	-4.252	.000	-1.468	-.534	-.246	-.393	-.318
CYP2C9*2_3	-.603	.195	-.237	-3.095	.003	-.989	-.216	-.174	-.297	-.232
CYP2C9*2_2	-.256	.089	-.229	-2.880	.005	-.433	-.080	-.070	-.278	-.216
CYP2C9*3_2	-.264	.087	-.236	-3.042	.003	-.437	-.092	-.090	-.292	-.228
VKORC1 - 1639G>A-HETERO	-.177	.078	-.211	-2.287	.024	-.331	-.023	.117	-.224	-.171
X1(AGE)	-.048	.023	-.157	-2.058	.042	-.094	-.002	-.207	-.203	-.154

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
.667	.445	.406	.32452	.024	4.236	1	99	.042

The results of MLE applied on the curated Turkish data set for 5 factors (Model-4). The dependent Variable Y is Warfarin Dose.
a.Unstandardized Coefficients
b.Standardized Coefficients
c.Intercept

4.4.3 Selecting Priors

If there is reliable background information, it will be reasonable to center the prior distributions for the mean and standard deviation of the variables. By this way, each version of a model is not starting from scratch, based only on the present data, but the cumulative effects of all data related to the previous studies can be taken into account. In the absence of reliable background information, non-informative priors are used. A non-informative (vague) prior is a conjugate prior with a large scale parameter. However, non-informative priors can pose problems when the sample size is small. Therefore, Weakly Informative Prior (WIP) distributions are utilized which make use of prior information for regularization and stabilization, providing enough prior information to prevent results that contradict our knowledge or problems such as negative variance. Moreover, no prior distributions are ever completely non-informative, not even a uniform distribution over the entire range of allowable values, and because it would cease to be uniform if the parameter were transformed.

Categorical and binary variables do not conform to normality principle. This type of highly skewed discrete data is dealt with a threshold approach where the discrete data

is treated as manifestations of an underlying normal distribution. In AMOS, it is assumed that there is an underlying continuous numerical variable whose range of values (minus infinity to plus infinity) is divided up into non-overlapping intervals. Therefore, recoded binary and categorical variables are treated like continuous variables in-terms of assessing the prior distributions.

In AMOS, one of the following families of prior distributions can be chosen for an individual model parameter:

- Uniform: You can specify the lower and upper bounds.
- Normal: You can specify the mean and standard deviation.
- Custom: You can make a free-hand sketch of the distribution after specifying its lower and upper bounds

In the initial phases of the study, the uninformative priors were used for all of the parameters of IWPC data set which was relatively a large sample size (after pre-processing ≈2500 rows were used in the training cohort) judging that the influence of the prior distribution would diminish as the sample size increased. AMOS uses diffuse priors and applies a uniform distribution from -3.4×10^{-38} to 3.4×10^{38} to each parameter. Due to poor convergence, improper solutions (negative variance estimates were observed) and crashes experienced during estimation, weakly-informative priors and informative priors were started to be used for the continuous observed variables based on the reference studies. Yet another reason for using informative priors was the use of Turkish data set. The sample size should be large enough to ensure the reliability of the parameter estimates. It was suggested that a SEM containing latent variables should have a minimum sample size of 200. The usage of priors was also indicated to overcome this shortcoming [110], [111].

Consequently, the constraints on Prior distributions were set for variables on a parameter-by-parameter basis by taking into account the academic literature and elicited information from the domain experts. Note also that, since a communized set of factors were used for both the IWPC and Turkish data set, the same type prior of distributions with similar settings were applied for the corresponding models. The prior distributions of the means of the coefficients of the exogenous observed (manifest) variables were set to a bounded uniform distribution indicating the influence of the factor on the warfarin dosing. For instance, Age is indicated to be a factor associated with a reduced dose of warfarin because the risk of warfarin triggered bleedings increases with advancing age [11], [24], [69]. Thus, the mean of the coefficient of Age was assigned to uniform distribution having a lower bound of -1 and upper bound of 0. On the other hand, BSA is associated with a higher dose of warfarin and therefore the mean of corresponding coefficient of BSA variable in Model-1 and Model-3 was assigned to a uniform prior distribution having 0 as the lower bound and 1 as the upper bound. Genetic factors such as VKORC1 -1639G>A, CYP2C9*2 and CYP2C9*3 are associated with lower dose requirements and the prior distribution of their coefficient means were assigned to a uniform distribution having -1 as the lower bound and 0 as the upper bound [22], [24], [26], [33]. The SNPs of CYP4F2, which is included in the Turkish data set, is indicated with a higher dose requirement and thus the prior distribution of the corresponding coefficient mean was assigned to a uniform distribution having 0 as the lower bound

and 1 as the upper bound [21]. These bounded uniform distributions are regarded as proper Weakly Informative Priors (WIP) [14], [20], [110], [111].

All of the prior distributions of the variances of the observed exogenous variable were assigned to WIPs with a uniform distribution having 0 as lower the bound and 3.4×10^{38} (AMOS default) as upper bound. These improper priors were re-organized using the Admissibility test if needed to avoid improper posteriors. The priors of other variables were set to AMOS default as expressed above.

The assessment of the influence imposed on the posterior estimates by prior distributions is also very crucial to ensure that a proper solution is obtained. The sensitivity of the prior assumptions was evaluated by repeating the analysis under different prior assumptions and comparing the posterior results for each unknown. As the size of a dataset grows, the evidence from the data eventually disregards the priori information, and the influence of the prior distribution decreases. To verify this assumption, Bayesian inference was re-performed after the prior distributions for the subjected parameter were reset to default (uniform distribution having a lower bound of -3.4×10^{-38} and an upper bound 3.4×10^{38}). For instance, the priors of the coefficient of Age were set to default and almost the same Mean and Standard Error estimates (-0.026 and 0.002) were obtained indicating the sensitivity posteriors to priors is low for Model-3 of IWPC dataset. The same sensitivity check was also applied for the Turkish data set, and a small discrepancy was observed in the posterior Mean and Standard Error of Age (-0.024 and 0.013) meaning that priori information had relatively higher influence on the posterior estimate. This finding is consistent with the theoretical assertion which states that small data sets are more sensitive to the usage of informative priors and as the size of the sample increases the effect of the priors diminishes [20], [110].

4.4.4 Defining analysis properties

In AMOS, MCMC properties are changed from the menu by choosing View – Options. Then, the MCMC tab is selected in the Options dialog box, A burn-in value of 1000 (default 500) and a refreshing interval (default 1000) of 2000 were selected. The number of MCMC iterations was set to 200,000.

4.4.5 Testing for convergence

The convergence can be checked by several diagnostics in AMOS. The built-in Gelman-Rubin diagnostic for convergence was used [39], [88], [110]. The specified models were by default accepted to have been converged when the diagnostic values were less than 1.002. Note that, considering that the MCMC chain has converged by this criterion does not mean that the summary table will stop changing. It is suggested that the summary table should be inspected as soon as the estimates continue to change and errors continue to decrease.

4.4.6 Goodness of Fit and Model selection

In this study, the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI) were used to test for model fit of a single model. For instance, the RMSEA was calculated as 0.057 (values about 0.05 or less are accepted as good fit) and the CFI was reported as 0.92 (a value of CFI $\geq$ 0.95 is recognized as an indication of good fit) for the best fitted candidate model of the Model-3 [56].

For comparing two non-nested models, the Browne-Cudeck Criterion (BCC) and the Akaike Information Criterion (AIC) value were utilized. Among the candidate fitted models for the Model-3, the model with the rescaled BCC value of 1.642 and the rescaled AIC value of 1.723 was selected [54].

CHAPTER 5

RESULTS AND DISCUSSION

This chapter outlines the results of Bayesian based modelling and estimation for warfarin Dosing and also discusses these results with respect to the previous studies.

Table 13: The Results of Bayesian Based Modelling and Estimation for Warfarin Dosing in IWPC and Turkish Data Sets (Model-1 and Model-2)

Parameter Estimates (Bayesian Estimation for IWPC Data Set)				
			95% Confidence Interval	
Parameter	Mean	Std. Error	Lower Bound	Upper Bound
b0 (intercept)	.298	.022	.265	.324
Age	-.034	.002	-.038	-.029
CYP2C9*2_2	-.051	.019	-.082	-.025
CYP2C9*3_2	-.122	.013	-.136	-.109
VKORC1 -1639G>A-MUTANT	-.277	.009	-.288	-.265
CYP2C9*2_3	-.102	.009	-.115	-.090
CYP2C9*3_3	-.205	.028	-.245	-.167
VKORC1 -1639G>A-HETERO	-.105	.015	-.121	-.090
BSA	.156	.012	.132	.180
Amiodarone	-.103	.014	-.130	-.076
DVT/PE	.009	.002	.007	.011
Smoking	.019	.005	.012	.026
Target-INR	.095	.014	.078	.112

Parameter Estimates (Bayesian Estimation for Turkish Data Set)				
			95% Confidence Interval	
Parameter	Mean	Std. Error	Lower Bound	Upper Bound
b0 (intercept)	3.823	.383	3.062	4.583
Age	-.054	.024	-.101	-.007
CYP2C9*2_2	-.270	.091	-.451	-.089
CYP2C9*2_3	-.594	.198	-.988	-.200
CYP2C9*3_2	-.264	.087	-.437	-.091
CYP2C9*3_3	-.976	.235	-1.442	-.510
VKORC1 -1639G>A-HETERO	-.189	.079	-.345	-.032
VKORC1 -1639G>A-MUTANT	-.657	.098	-.851	-.463
CYP4F2-HETERO	.091	.028	.054	.138
CYP4F2-MUTANT	.210	.096	.020	.401
BSA	.130	.014	.088	.157
Indications	.028	.009	.011	.043

The Model-1 and the Model-2 were provided for a complete eligible set of variables. The results of the corresponding stepwise regressions for the Model-1 and the Model-2 were consistent with the referenced linear dosing algorithms [2],[4],[55].The results of the Model-1 and the Model-2 established by Bayesian SEM are provided in Table 13.

On the other hand, the Model-3 and the Model-4 were established for comparison and validation purposes. Exactly the same set of features from entirely different data sets were modelled in a similar fashion and Bayesian inference was performed.

MLE explained 38.1% to 46.3% of the variation (R^2) in warfarin dose for IWPC data set using 5 factors (Model-3) and 10 factors (Model-1) respectively. On the other hand, MLE could explain 44.5% of the variation (R^2) in warfarin dose for Turkish data set using 5 factors (Model-4). The performances of the linear regression algorithms were calculated as 45.1% and 49.3% for Model-3 and Model-4. For the same 5 factors, the proportion of variation explained (R^2) by the corresponding linear model was higher for the small Turkish dataset than the large IWPC dataset. In addition to that, the linear model applied on Model-4 also performed better than Model-3 supposedly owing to the diverse ethnic population of the IWPC dataset and the data size. It was also assessed that the standard deviation of warfarin dose in Turkish data set was higher than the standard deviation of warfarin dose in IWPC data set which caused a higher percentage explained by Bayesian estimation for the Turkish data set. The Bayesian modelling and estimation was initially applied for both of the training cohorts utilizing the 5 common factors and the resultant estimates of the best-fitted models were recorded (Table 14).

Table 14: The Results of Bayesian based Modelling and Estimation for Warfarin Dosing in IWPC and Turkish Data Sets (Model-3 and Model-4)

Parameter Estimates

(Bayesian Estimation for IWPC Data Set)

Parameter	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
b0 (intercept)	.412	.019	.374	.450
Age	-.025	.002	-.028	-.022
CYP2C9*2_2	-.046	.021	-.087	-.006
CYP2C9*3_2	-.091	.005	-.101	-.081
VKORC1 -1639G>A-MUTANT	-.177	.007	-.192	-.163
CYP2C9*2_3	-.088	.007	-.101	-.075
CYP2C9*3_3	-.173	.028	-.229	-.118
VKORC1 -1639G>A-HETERO	-.077	.009	-.095	-.059

Parameter Estimates

(Bayesian Estimation for Turkish Data Set)

Parameter	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
b0 (intercept)	1.152	.049	1.051	1.284
Age	-.020	.009	-.038	-.001
CYP2C9*2_2	-.098	.036	-.169	-.027
CYP2C9*2_3	-.256	.043	-.291	-.212
CYP2C9*3_2	-.100	.035	-.170	-.030
CYP2C9*3_3	-.442	.032	-.404	-.479
VKORC1 -1639G>A-HETERO	-.063	.029	-.121	-.004
VKORC1 -1639G>A-MUTANT	-.267	.040	-.346	-.189

The results of Bayesian Estimation applied for both IWPC (Model-3) and Turkish (Model-4) data sets for 5 common factors (CYP2C9*2, CYP2C9*3, and VKORC1 -1639G>A are recoded to binary variables beforehand).

Based on the same genetic and non-genetic factors, Bayesian Estimation was indicated to explain the 44.6% (38.1% by MLE) and 51.2% (44.5% by MLE) of the variation (R^2) in dosing for IWPC and Turkish data sets respectively. Note that although the Bayesian estimation improved the portion of the variance explained for

both data sets, it was more successful for the Turkish data set for 5 factors (Model-4). The prediction performances were also improved for both data sets (47.4% and 51.7% respectively) compared to MLE (45.1% and 49.3%).

Bayesian estimations were repeated for the randomly formed test cohorts of the two data sets as well. The results obtained for the best fitted models of each data set were compared to the results obtained from training cohorts. It was concluded that the results were statistically consistent in the 95% confidence interval ($p = 0.0023$).

Keeping the 5 dominant factors in the model, new factors were separately introduced into Bayesian SEMs for each data set incrementally. For the model that includes the complete set of factors from Turkish data set (Model-2), VKORC1:-1639G>A was the most dominant genetic feature and explained 27.2% of the variance in warfarin dosing (Mutant: 19.8%, Hetero: 7.4%). The second significant gene was CYP2C9. CYP2C9*3 explained 13.2% of the variance in maintenance dose, while CYP2C9*2 explained 10.4%. CYP4F2 was associated with 3.2% of the variance. None-genetic factors were also associated with warfarin dose. The factors age, BSA, and indications (PE/DVT, Cardiac Indications) were also accounted for warfarin dose (2.8%, 1.2%, 0.7%). Thus, model including 8 factors from the Turkish data set could explain 56.7% of the variance.

As a second task the model for IWPC data set was also enriched with the new factors available. For IWPC data set, BSA, PE/DVT, amiodarone, target-INR and smoking were incorporated (Model-1). The total variance explained for IWPC data set using Bayesian Estimation was found to be 53.9%. We studied the case when the genetic varıables were treated as continuous variables on the IWPC data set. (For CYP2C9, wild, heterozygous, and mutant genotypes were accepted as 0, 1, and 2 respectively). The results have shown that the percentage of the variance explained dropped from 53.9% to 50.6%.

For several reasons, warfarin is an ideal drug to test the hypothesis that pharmacogenomics can reduce drug toxicity: it is commonly prescribed, has a narrow therapeutic/toxic ratio, and is affected by common genetic polymorphisms. The results of the pharmacogenomics studies caused the U.S. Food and Drug Administration (FDA) to update the warfarin drug label to include information about genetic polymorphisms affecting dosing in August 2007 [13]. The FDA modified the warfarin label, stating that VKORC1 and CYP2C9, age, BSA, interacting drugs, and indication for warfarin therapy explained about 55% of the variability in warfarin dose in Caucasian patients. These results were supported by the IWPC's rather complicated equation which had been provided in the supplementary appendix of Klein et al. [11].

The previous studies which proposed pharmacogenomics driven formulations for the prediction of warfarin dosage used a model based on linear regression [11], [12], [22] [24]. Roper et al. [25] conducted a comparative study to validate and compare the three published warfarin dosing algorithms by Sconce et al. [12], Anderson et al. [24], Klein et al. [11] and the *WarfarinDosing* website [80] based on the studies of Gage et al. [22], [23]. The predicted dose requirements were compared with the actual maintenance dose for each patient within the therapeutic INR of 2.0 to 3.0. It

was concluded that all linear regression based methods produced similar results explaining only 37.7% to 45.8% of the variance in warfarin dosage in the IWPC training data set as indicated in Table V. It was also shown that when comparing the percentage of patients whose predicted dosage were within 20% of actual, the IWPC algorithm performed the best (45.9%).

The distribution of the Bayesian estimations obtained for target variables is compared to the distribution of the actual clinical values. In the 95% confidence interval, p is calculated as 0.0048 ($p < 0.005$) meaning that the Bayesian model gives the same results as of the clinical ones. A second test is performed to assess the influence of the pharmacogenomics variables. The distribution obtained from the Bayesian SEM including the pharmacogenomics variables is compared to the distribution obtained from the Bayesian SEM without the pharmacogenomics variables. In the 95% confidence interval, p is calculated as 0.0063 ($p > 0.005$) meaning that the pharmacogenomics variables are effective for explaining the variance in warfarin dosage.

Bayesian SEM is also utilized without pharmacogenomics variables based on the IWPC data set. The model that included BSA, age, amiodarone, Target-INR, DVT/PE as the factors converged and explained the 21.7% of the variability in dosing. The results are consistent with the literature information indicating that 17−25% of the variability in the therapeutic warfarin dose can be explained by using clinical factors alone [22]. Thus, Bayesian SEM can also be effectively used for clinical factors alone.

CHAPTER 6

CONCLUSIONS

This chapter summarizes the study and its findings, describes the limitations and suggests possibilities for further research.

6.1 Summary

The incorporation of pharmacogenomics information into the drug dosing estimation formulations has been shown to increase the accuracy in drug dosing and decrease the frequency of adverse drug effects in many studies in the literature [6], [8], [9]. Extensive pharmacogenomics research efforts have identified several genetic factors that are strongly associated with drug dosing such as warfarin pharmacogenomics and the effect of CYP2D6 on tamoxifen efficacy. However, these findings are mainly expressed as associations rather than predictors as validated practical tools. Besides, these efforts are not consistent in determining the criteria for the accuracy of dosing algorithms [25]. Therefore a robust estimation framework is needed to assess these interactions thoroughly using a holistic approach while facilitating accurate dosing predictions. Such an estimation framework -regardless of the drug used- should address the following issues [36], [73]:

a) Clinical validity: Is a pharmacogenomics factor associated with an adverse effect indicated to be a clinically useful predictor for that adverse effect?

b) Clinical utility: Does the incorporation of the genetic factor predict the outcome more accurately than existing clinical models?

c) Degree of clinical utility: Are the predictions for indicated to be sufficiently different to change treatment rulings?

Warfarin is an excellent drug to investigate such an estimation framework for drug dosing. First of all, it is placed in the “top 10 drugs” for ADR related hospitalizations in the US [62]. Secondly, there is a considerable individual variation in warfarin response due to primarily genetic factors. Thirdly, it is a commonly used anti-coagulant drug having a narrow therapeutic window (i.e., the under-dosing and over-dosing of warfarin can be disastrous due to the thrombotic or hemorrhagic adverse reactions respectively). Lastly, the initiation of warfarin therapy based on clinical procedures is risky and problematic. Thus, various algorithms for the estimation of

warfarin dosing have been proposed (e.g., the warfarin dose calculator provided by IWPC - PharmGKB) [11], [22].

CYP2C9 and VKORC1 genotypes are clinically useful predictors of warfarin dose in clinical trials. As demonstrated by multiple studies, including the work of IWPC [11], dosing based on clinical/demographic factors alone improves prediction of stable therapeutic dose of warfarin (compared to the one-size-fits-all 5 mg/day dose). Moreover, the FDA modified the warfarin label, stating that VKORC1 and CYP2C9, age, BSA, interacting drugs, and indication for warfarin therapy explained about 55% of the variability in warfarin dose. Besides, the incorporation of CYP2C9 and VKORC1 genotypes improves prediction of warfarin doses as confirmed by randomized clinical trials proving effective utilization of pharmacogenomics.

This study which aims to establish a robust pharmacogenomics driven modelling and estimation framework for warfarin dosing, fits well with the main objective: To take a further step in the direction of personalized medicine. It has been shown that the Bayesian modelling and estimation framework for pharmacogenomics driven drug dosing better explains the variations in warfarin dosing and improves prediction of stable therapeutic dose of warfarin compared to the linear estimation algorithms. The main genetic (VKORC1 -1639G>A, CYP2C9*2, CYP2C9*3) and non-genetic factors (Age, BSA, Amiodarone, Target-INR) are found to be consistent with the previous studies and literature.

In conclusion, the results indicate that the proposed framework based on Bayesian SEM better explains the variation in warfarin dosing and improves the prediction accuracy compared to the state-of-the-art methods. The proportion of variation explained (R^2) was higher (44.6%-56.7%) depending on the data set and the features included in the model. The predicted dose is compared to the actual dose for each patient to assess the performance of the dosing algorithm. The proposed framework including 5 dominant factors also performs better than the IWPC algorithm (47.4% vs 45.1%) given that the predicted doses were within 20% of the actual. However, it is strongly suggested that the model should be extensively tested and validated in various clinical settings.

It has been also indicated that Bayesian SEM is a robust and effective approach for the estimation of drug dosing since it facilitates the exploration and identification of hidden relationships and provides the flexibility to utilize useful prior information for achieving better prediction results [112]. However, the prior information and data should be dealt with carefully so as to achieve reliable unbiased predictions. Besides Bayesian sampling does not rely on asymptotic theory and thus give more reliable results for small data sets (e.g. Turkish data set). Therefore, Bayesian SEM may be very practical and effective for clinical studies where the data acquisition may be problematic and expensive. On the other hand, Bayesian estimation often gives results that are close to the results provided by frequentist methods when the sample size increases.

Similar studies [82], [92], [113] in other subject areas of Medical Informatics also suggest that Bayesian modeling and estimation can be a further step in the direction of personalized medicine. Nevertheless, the future of personalized warfarin dosing

depends on the availability of easily applicable and in-expensive genotyping methods.

6.2 Limitations

The limitations to this study generally arise from the fact that Bayesian SEM is mainly a data driven method. The study is limited to data sets worked upon and the features that are being missing or a portion of the data inhibits or degrades the Bayesian estimation. Although Bayesian SEM has the capability to handle missing data to some extent, it is limited with the availability of data at the end of the day. For instance, some of the features related with warfarin dosing such as medications, and herbal and nutritional intakes (vitamin K) cannot be included into the model although they exist in the hypothetical model due to the lack of data. Similarly, some pharmacogenomics features that exist in one study and are not used in another study (e.g., CYP4F2, EPHX1) due to pricing introduced by genotypic methods. Such an inadequacy of genetic data can cause problems for testing the validation and accuracy of the models.

Other limitations are related with the technical features of the Bayesian SEM. First, the prior distributions have to be selected carefully since the prediction accuracy of Bayesian SEM relies on the priors. If the priors are not set correctly, then misleading results will be generated. Besides, there is no exact method for choosing a prior. Therefore, Bayesian SEM requires skills and knowledge to convert prior beliefs into mathematically formulated priors. Second, Bayesian estimation often incurs a high computational cost, especially for models with a large number of parameters, which may be time-consuming and cumbersome.

6.3 Future Work

Although this study proposed a hypothetical framework involving all the related factors of warfarin dosing, several factors that are literally indicated to be accounted for variations in warfarin dosing could not be studied due to the lack of data. Especially the absence of pharmacogenomics data affects the performance of the Bayesian estimation. Therefore, as a major future work, a collaborative clinical study can be carried out by multiple medical centers so as to collect diversified but complete data consisting of a set of pre-defined factors. Such an extensive collaborative study can cover the factors which have not been utilized in this study. For instance, intakes of K vitamin and other nutrients can be very effective on warfarin dose and can be incorporated in the model. Co-morbidities, other medications, tea/coffee consumption and other uncovered genetic factors which can theoretically explain the variation in dosing can also be investigated within the scope of this collaborative study. It is likely that the framework would explain a greater percentage of the variation in dosing if such complete data were provided. Turkish data set [21] contains such information so that it can be worked upon in more detail to investigate the effect of K vitamin on warfarin dosing as a future work.

Despite the approval of CYP2C9/VKORC1 genotyping platforms by the U.S. FDA, the clinical implementation of genotype-guided dosing has been delaying over the last decade. Although genotype-guided therapy improves dose prediction, there has been limited number of studies to assess the effects (the amount of reduced risk with respect to ADRs and saving for health-care costs) of such a genotype-guided dosing. As another future work, studies that assess the results and effects of genotype-guided therapy can be carried out.

On the other hand, Bayesian SEM approach can also be applied for other drugs such as statins, selective serotonin reuptake inhibitors (SSRIs) and tamoxifen, which are affected by genetic and non-genetic factors.

The application of Bayesian approach to the Artificial Neural Networks (ANN) for the prediction drug dosing can be a promising future study. Bayesian network inference algorithms can capture linear, non-linear, combinatorial, stochastic and other types of relationships among variables.

REFERENCES

[1] WHO Patient Safety Research Advisory Council, "WHO Patient Safety Research," 2009.

[2] WHO, "International Drug Monitoring: The Role of International Centers," WHO, Geneva, 1972.

[3] A. P. Fletcher, "Spontaneous adverse drug reaction reporting vs event monitoring: a comparison," Journal of the Royal Society of Medicine, vol. 84, pp. 341-344, 1991.

[4] B. Begaud, Y. Moride, P. Tubert-Bitter, A. Chaslerie and F. Haramburu, "False-positives in spontaneous reporting: should we worry about them?," British Journal of Clinical Pharmacology, vol. 38, pp. 401-404, 1994.

[5] I. S. Yun, M. J. Koo, H. E. Park, S.-E. Kim, J.-H. Lee, J.-W. Park and C.-S. Hong, "A Comparison of Active Surveillance Programs Including a Spontaneous Reporting Model for Phamacovigilance of Adverse Drug Events in a Hospital," Korean J Intern Med, vol. 27, pp. 443-450, 2012.

[6] A. L. Beitelshees and H. L. McLeod, "Applying pharmacogenomics to enhance the use of biomarkers for drug effect and drug safety," TRENDS in Pharmacological Sciences, vol. 27, no. 9, pp. 498-502, 2006.

[7] E. Krynetskiy and P. McDonnell, "Building individualized medicine: prevention of adverse reactions to warfarin therapy," J Pharmacol Exp Ther, vol. 322, no. 2, pp. 427-434, 08 2007.

[8] L. Becquemont, "Pharmacogenomics of adverse drug reactions: practical applications and perspectives," Pharmacogenomics, vol. 10, no. 6, pp. 961-969, 06 2009.

[9] D. W. Clark, E. Donnelly, D. M. Coulter, R. L. Roberts and M. A. Kennedy, "Linking Pharmacovigilance with Pharmacogenetics," Drug Safety, vol. 27, no. 15, pp. 1171-1184, 2004.

[10] K. J. Karczewski, R. Daneshjou and R. B. Altman, "Chapter 7: Pharmacogenomics," PLOS Computational Biology, vol. 8, no. 12, pp. 1-18, 2012.

[11] IWPC, T. E. Klein, R. B. Altman, N. Eriksson, B. F. Gage, S. E. Kimmel, M. T. Lee, N. A. Limdi, D. Page, D. M. Roden, M. J. Wagner, M. D. Caldwell and J. A. Johnson, "Estimation of the warfarin dose with clinical and pharmacogenetic data," New England Journal of Medicine, vol. 360, no. 8, pp. 753-764, 2009.

[12] E. A. Sconce, T. I. Khan, H. A. Wynne, P. Avery, L. Monkhouse, B. P. King, P. Wood, P. Kesteven, A. K. Daly and F. Kamali, "The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen," Blood, vol. 106, no. 7, p. pp. 2329–2333, 2005.

[13] U.S. Food and Drug Association, "News And Events," 16 08 2007. [Online]. Available: http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2007/ucm108967.htm. [Accessed 11 06 2014].

[14] S.-Y. Lee, Structural Equation Modeling: A Bayesian Approach, West Sussex: John Wiley and Sons Ltd, 2007.

[15] P. Armitage, G. Berry and J. N. S. Matthews, Statistical Methods in Medical Research, Cornwall: Blackwell Publishing, 2002.

[16] StataCorp LP, "STATA Structural Equation Modelling Reference Manual Release 13," College Station, 2013.

[17] J. Grace, "Modeling with Structural Equations," 2012. [Online]. Available: http://www.structuralequations.org/. [Accessed 11 06 2014].

[18] G. D. Garson, Structural Equation Modeling, Statistical Associates Publishers, 2012.

[19] D. Gefen, D. Straub and M.-C. Boudreau, "Structural Equation Modeling and Regression: Guidelines for Research Practice," Communications of the Association for Information Systems, vol. 4, pp. 1-79, 2000.

[20] D. B. Dunson, J. Palomo and K. Bollen, "Bayesian Structural Equation Modeling," 27 Temmuz 2005. [Online]. Available: http://www.samsi.info/communications/bayesian-structural-equation-modeling. [Accessed 11 06 2014].

[21] M. Özer, Y. Demirci, C. Hızel, S. Sarıkaya, İ. Karaltı, Ç. Kaspar, S. Alpan and E. Genç, "Impact of Genetic Factors (CYP2C9, VKORC1 and CYP4F2) on Warfarin Dose Requirement in the Turkish Population," Basic and Clinical Pharmacology & Toxicology, vol. 112, no. 3, pp. 209-214, March 2013.

[22] B. F. Gage, C. Eby, J. A. Johnson, E. Deych, M. J. Rieder, P. M. Ridker, P. E. Milligan, G. Grice, P. Lenzini, 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, "Use of Pharmacogenetic and Clinical Factors to Predict the Therapeutic Dose of Warfarin," Clinical Pharmacology and Therapeutics, vol. 84, no. 3, pp. 326-331, 2008.

[23] E. C. M. P. B. G. D. J. M. H. Gage BF, "Use of pharmacogenetics and clinical factors to predict the maintenance dose of warfarin," Thrombosis and Haemostasis, vol. 91, no. 1, pp. 87-94, 2004.

[24] H. B. S. S. W. S. S. K. M. J. R. M. B. S. B. K. M. C. H. J. R. J. S. D. B. T. K. S. M. J. C. J. Anderson JL, "A randomized and clinical effectiveness trial comparing two pharmacogenetic algorithms and standard care for individualizing warfarin dosing (coumagen-II)," Circulation, vol. 125, pp. 1997-2005, 2012.

[25] N. Roper, B. Storer, R. Bona and M. Fang, "Validation and Comparison of Pharmacogenetics-Based Warfarin Dosing Algorithms for Application of Pharmacogenetic Testing," Journal of Molecular Diagnostics, vol. 12, no. 3, pp. 283-291, 2010.

[26] D. E. Jonas and H. L. McLeod, "Genetic and clinical factors relating to warfarin dosing," Trends in Pharmacological Sciences, vol. 30, no. 7, pp. 375-386, 2009.

[27] B. S. Finkelman, B. F. Gage, J. A. Johnson, C. M. Brensinger and S. E. Kimmel, "Genetic warfarin dosing: Tables versus Algorithms," J Am Coll Cardiol., vol. 52, no. 5, pp. 612-618, 1 2 2011.

[28] M. Wadelius and M. Pirmohamed, "Pharmacogenetics of Warfarin: Current Status and Future Challenges," The Pharmacogenomics Journal, vol. 7, no. 2, p. 99–111, 2007.

[29] PharmGKB - IWPC, "IWPC - International Warfarin Pharmacogenetics Consortium," 2008. [Online]. Available: http://www.pharmgkb.org/page/iwpc. [Accessed 11 06 2014].

[30] R. P. Owen, R. B. Altman and T. E. Klein, "PharmGKB and the International Warfarin Pharmacogenetics Consortium: The Changing Role for Pharmacogenomic Databases and Single-Drug Pharmacogenetics," Human Mutation, vol. 29, no. 4, pp. 456-460, 2008.

[31] K. Sangkuhl, D. S. Berlin, R. B. Altman and T. E. Klein, "PharmGKB: Understanding the Effects of Individual Genetic Variants," Drug Metab Review, vol. 40, no. 4, pp. 539-551, 2008.

[32] F. Takeuchi, R. McGinnis, S. Bourgeois, C. Barnes, N. Eriksson, N. Soranzo, P. Whittaker, V. Ranganath, V. Kumanduri, W. McLaren, L. Holm, J. Lindh, A. Rane, M. Wadelius and P. Deloukas, "A Genome-Wide Association Study Confirms VKORC1, CYP2C9, and CYP4F2 as Principal Genetic Determinants of Warfarin Dose," PLoS Genetics, vol. 5, no. 3, pp. 1-9, 2009.

[33] J. Lindh, "Major Determinants of Outcome and Dosing in Warfarin Treatment," Karolinska Institutet, Stockholm, 2009.

[34] C. Sanoski and J. Bauman, "Clinical observations with the amiodarone/warfarin interaction: dosing relationships with long-term therapy," Chest, vol. 121, pp. 19-23, 2002.

[35] N. A. Limdi, M. Wadelius, L. Cavallari, N. Eriksson, D. C. Crawford, M.-T. M. Lee, C.-H. Chen, A. Motsinger-Reif, H. Sagreiya, N. Liu, A. H.-B. Wu, B. F. Gage, A. Jorgensen, M. Pirmohamed, J.-G. Shin, G. Suarez-Kurtz, S. E. Kimmel, J. A. Johnson, T. E. Klein and M. J. Wagner, "Warfarin pharmacogenetics: a single VKORC1 polymorphism is predictive of dose across 3 racial groups," Blood, vol. 115, no. 18, p. 3827–3834, 06 05 2010.

[36] N. A. Limdi, "Warfarin pharmacogenetics: challenges and opportunities for clinical translation," Frontiers in Pharmacology, pp. 1-5, 17 10 2012.

[37] D. J. Greenblatt and L. L. von Moltke, "Interaction of Warfarin With Drugs, Natural Substances, and Foods," Journal of Clinical Pharmacolology, vol. 45, no. 2, pp. 127-132, 02 2005.

[38] Y. Lurie, R. Loebstein, D. Kurnik, S. Almog and H. Halkin, "Warfarin and vitamin K intake in the era of pharmacogenetics," Br J Clin Pharmacol., vol. 70, no. 2, pp. 164-70, 08 2010.

[39] J. L. Arbuckle, "IBM® SPSS® Amos™ 21 User Guide," [Online]. Available: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/amos/21.0/en/Manuals/IBM_SPSS_Amos_Users_Guide.pdf. [Accessed 15 09 2013].

[40] E. C. Davies, C. F. Green, S. Taylor, P. R. Williamson, D. R. Mottram and M. Pirmohamed, "Adverse Drug Reactions in Hospital In-Patients: A Prospective Analysis of 3695 Patient-Episodes," PLoS ONE, vol. 4, no. 2, pp. 1-7, 11 02 2009.

[41] J. Lazarou, B. H. Pomeranz and P. N. Corey, "Incidence of Adverse Drug Reactions in Hospitalized Patients," The Journal of The American Medical Association, vol. 279, no. 15, pp. 1200-1205, 15 04 1998.

[42] H. J. Murff, V. L. Patel and G. Hripcsak, "Detecting adverse events for patient safety research: a review of current methodologies," Journal of Biomedical Informatics, vol. 36, no. 1-2, p. 131–143, 2003.

[43] B. H. Stricker and B. M. Psaty, "Detection, verification, and quantification of adverse drug reactions," British Medical Journal,, vol. 329, no. 1, pp. 44-47, 2004.
[44] WHO, 10 2004. [Online]. Available: http://apps.who.int/medicinedocs/pdf/s6164e/s6164e.pdf. [Accessed 11 06 2014].

[45] L. Mancinelli, M. Cronin and W. Sadée, "Pharmacogenomics: The Promise of Personalized Medicine," AAPS PharmSci, vol. 2, no. 1, pp. 29-41, 03 2000.

[46] WHO Collaborating Centre for International Drug Monitoring, "The Importance of Pharmacovigilance," Geneva, 2002.

[47] K. C. Frazier, "The lessons of Vioxx," The New England Journal of Medicine, vol. 353, pp. 2576-8, 09 2005.

[48] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, vol. 17, no. 3, pp. 37-54, 1996.

[49] A. M. Wilson, L. Thabane and A. Holbrook, "Application of data mining

techniques in pharmacovigilance," British Journal of Clinical Pharmacology, vol. 57, no. 2, pp. 127-134, 02 2004.

[50] G. Hripcsak, S. Bakken, P. D. Stetson and V. L. Patel, "Mining complex clinical data for patient safety research: a framework for event discovery," Journal of Biomedical Informatics, vol. 36, no. 1-2, pp. 120-130, 2003.

[51] D. M. Fram, J. S. Almenoff and W. DuMouchel, "Empirical Bayesian data mining for discovering patterns in post-marketing drug safety," in ACM SIGKDD international conference on Knowledge discovery and data mining, New York, 2003.

[52] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner and R. M. De Freitas, "A Bayesian neural network method for adverse drug reaction signal generation," Eur J Clin Pharmacol, vol. 54, no. 4, pp. 315-321, 1998.

[53] A. Bate, "The Use of a Bayesian Confidence Propagation Neural Network in Pharmacovigilance," Umeå University, Umeå, 2003.

[54] M. Lindquist, M. Ståhl, A. Bate A, I. R. Edwards and R. H. Meyboom, "A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database," Drug Safety, vol. 6, pp. 533-42, 12 2000.

[55] C. Bousquet, C. Henegar, A. L.-L. Louët, P. Degoulet and M.-C. Jaulent, "Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach," International Journal of Medical Informatics, vol. 74, no. 7-8, p. 563—571, 2005.

[56] C.-L. Liew, J.-H. Yen and A.-B. Liu, "Gender differences in the effective warfarin dosage in Han and aboriginal Taiwanese patients with the VKORC1-1639AA genotype," Tzu Chi Medical Journal, vol. 25, no. 4, pp. 213-217, 2013.

[57] M. L. Becker and J. S. Leeder, "Identifying genomic and developmental causes of adverse drug reactions in children," Pharmacogenomics., vol. 11, no. 11, p. 1591–1602, 11 2010.

[58] J. Kirchheiner, K. Brøsen, M. L. Dahl, L. F. Gram, S. Kasper, I. Roots, F. Sjöqvist, E. Spina and J. Brockmöller, "CYP2D6 and CYP2C19 genotype-based dose recommendations for antidepressants: a first step towards subpopulation-specific dosages," Acta Psychiatr Scand, vol. 104, no. 3, pp. 173-192, 09 2001.

[59] A. P. Chiang and A. J. Butte, "Data-driven Methods to Discover Molecular Determinants of Serious Adverse Drug Events," Clinical Pharmacology and Therapeutics, vol. 85, no. 3, pp. 259-268, 03 2009.

[60] U.S. DoHHS, "Realizing the Potential of Pharmacogenomşics: Opportunities and Challenges," Washington, 2008.

[61] D. S. Budnitz, N. Shehab, S. R. Kegler and C. L. Richards, "Medication Use Leading to Emergency Department Visits for Adverse Drug Events in Older Adults," Annals of Internal Medicine, vol. 147, no. 11, pp. 755-65, 12 2007.

[62] D. S. Budnitz, M. C. Lovegrove, N. Shehab and C. L. Richards, "Emergency Hospitalizations for Adverse Drug Events in Older Americans," The New England Journal of Medicine, no. 365, pp. 2002-2012, 11 2011.

[63] C. F. Thorn, T. E. Klein and R. B. Altman, "Chapter 14: PharmGKB, The Pharmacogenetics and Pharmacogenomics Knowledge Base," in Pharmacogenomics, New Jersey, Humana Press, 2005, pp. 179-192.

[64] Indiana University, Division of Clinical Pharmacology, "P450 Drug Interaction Table," Indiana University, School of Medicine, Department of Medicine, 03 05 2014. [Online]. Available: http://medicine.iupui.edu/clinpharm/ddis/main-table/. [Accessed 11 06 2014].

[65] M. Ingelman-Sundberg, "Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future," TRENDS in Pharmacological Sciences, vol. 40, no. 4, pp. 193-200, 2004.

[66] M. G. Scordo, E. Aklillu, U. Yasar, M. L. Dahl, E. Spina and M. Ingelman-Sundberg, "Genetic polymorphism of cytochrome P450 2C9 in a Caucasian and a black African population," Br J Clin Pharmacol, vol. 52, no. 4, pp. 447-450, 2001.

[67] R. P. Owen, L. Gong, H. Sagreiya, T. E. Klein and R. B. Altman, "VKORC1 Pharmacogenomics Summary," Pharmacogenetics and Genomics, vol. 20, no. 10, p. 642–644, 10 2010.

[68] D. J. .. Harrington, R. Gorska, R. Wheeler, S. .. Davidson, S. .. Murden, C. Morse, M. J. .. Shearer and A. D. Mumford, "Pharmacodynamic resistance to warfarin is associated with nucleotide substitutions in VKORC1," Journal of Thrombosis and Haemostasis, vol. 6, no. 1, p. 1663–1670, 2008.

[69] V. L. Baczek, W. T. Chen, J. Kluger and C. I. Coleman, "Predictors of warfarin use in atrial fibrillation in the United States: a systematic review and meta-analysis," BMC Family Practice, vol. 13, no. 5, pp. 1471-2296, 2012.

[70] A. L. Jorgensen, R. J. FitzGerald, J. Oyee, M. Pirmohamed and P. R. Williamson, "Influence of CYP2C9 and VKORC1 on Patient Response to Warfarin: A Systematic Review and Meta-Analysis," PLoS ONE, 08 2012.

[71] C. Geisen, M. Watzka, K. Sittinger, M. Steffens, L. Daugela, E. Seifried, C. Müller, T. Wienker and J. Oldenburg, "VKORC1 haplotypes and their impact on the inter-individual and inter-ethnical variability of oral anticoagulation," Thrombosis and Haemostasis, vol. 94, no. 4, pp. 773-779, 2005.

[72] E. Pautas, C. Moreau, I. Gouin-Thibault, J. Golmard, I. Mahé, C. Legendre, E. Taillandier-Hériche, B. Durand-Gasselin, A. Houllier, P. Verrier, P. Beaune, M. Loriot and V. Siguret, "Genetic factors (VKORC1, CYP2C9, EPHX1, and CYP4F2) are predictor variables for warfarin response in very elderly, frail inpatients," Clin Pharmacol Ther., vol. 87, no. 1, pp. 57-64, 2010.

[73] N. A. Limdi and D. L. Veenstra, "Expectations, validity, and reality in pharmacogenetics," J Clin Epidemiol, vol. 63, no. 9, pp. 960-9, 09 2010.

[74] N. A. Limdi, T. M. Beasley, M. R. Crowley, J. A. Goldstein, M. J. Rieder, D. A. Flockhart, D. K. Arnett, R. T. Acton and N. Liu, "VKORC1 polymorphisms, haplotypes and haplotype groups on warfarin dose among African–Americans and European– Americans," Pharmacogenomics, vol. 9, no. 10, p. 1445–1458, 10 2008.

[75] M. D. Caldwell, T. Awad, J. A. Johnson, B. F. Gage, M. Falkowski, P. Gardina, J. Hubbard, Y. Turpaz, T. Y. Langaee, C. Eby, C. R. King, A. Brower, J. R. Schmelzer, I. Glurich, H. J. Vidaillet, S. H. Yale, K. Q. Zhang, R. L. Berg and J. K. Burmester, "CYP4F2 genetic variant alters required warfarin dose," Blood, vol. 111, no. 8, p. 4106–4112, 15 04 2008.

[76] J. Zhang, A. Jorgensen, A. Alfirevic, P. Williamson, C. Toh, B. Park and M. Pirmohamed, "Effects of CYP4F2 genetic polymorphisms and haplotypes on clinical outcomes in patients initiated on warfarin therapy," Pharmacogenet Genomics., vol. 19, no. 10, pp. 781-789, 2009.

[77] L. Harty, K. Johnson and A. Power, "Race and Ethnicity in the Era of

Emerging Pharmacogenomics," Journal of Clinical Pharmacology, vol. 46, no. 4, pp. 405-407, 2006.

[78] M. Lee, C. Chen, C. Chou, L. Lu, H. Chuang, Y. Chen, A. Saleem, M. Wen, J. Chen, J. Wu and Y. Chen, "Genetic determinants of warfarin dosing in the Han-Chinese population," Pharmacogenomics, vol. 10, no. 12, pp. 1905-1913, 2009.

[79] A. Liu and C. Stumpo, "Warfarin-Drug Interactions Among Older Adults," Geriatrics Aging, vol. 10, no. 10, pp. 643-646, 2007.

[80] B. F. Gage, "WarfarinDosing," Washington University in St. Louis, 05 02 2014. [Online]. Available: http://www.warfarindosing.org/Source/Home.aspx. [Accessed 11 06 2014].

[81] P. M. Bentler, A. Satorra and K.-H. Yuan, "Smoking and Cancers: Case-Robust Analysis of a Classic Data Set," Structural Equation Modeling: A Multidisciplinary Journal, vol. 16, no. 2, pp. 382-390, 2009.

[82] B. L. Fridley, G. Jenkins, D. J. Schaid and L. Wang, "A Bayesian hierarchical nonlinear model for assessing the association between genetic variation and drug cytotoxicity," Statistics in Medicine, vol. 28, no. 21, pp. 2709-2722, 20 09 2009.

[83] Z. Hongya, C. Kwok-Leung, C. Lee-Ming and Y. Hong, "Multivariate hierarchical Bayesian model for differential gene expression analysis in microarray experiments," BMC Bioinformatics, vol. 9, no. Suppl 1, pp. 1-10, 2008.

[84] MedLibrary.org, "Structural equation modeling," 2010. [Online]. Available: http://medlibrary.org/medwiki/Structural_equation_modeling. [Accessed 11 06 2014].

[85] G. D. Garson, "Testing Statistical Assumptions," Statistical Associates Publishing, 2012. [Online]. Available: http://www.statisticalassociates.com/assumptions.pdf.

[86] D. Hooper, J. Coughlan and M. R. Mullen, "Structural Equation Modelling: Guidelines for Determining Model Fit," Electronic Journal of Business Research Methods, vol. 6, no. 1, pp. 53-60, 2008.

[87] L. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," Structural Equation Modeling: A Multidisciplinary Journal, vol. 6, no. 1, pp. 1-55, 11 1999.

[88] A. Gelman, X.-L. Meng and . H. Stern, "Posterior predictive assessment of model fitness via realized discrepancies," Statistica Sinica, vol. 6, pp. 733-807, 1996.

[89] X.-Y. Song and S.-Y. Lee, "Bayesian Analysis of Structural Equation Models With Nonlinear Covariates and Latent Variables"," Multivariate Behavioral Research, vol. 41, no. 3, pp. 337-365, 2006.

[90] X. Wen and M. Stephens, "Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene-environment interactions," The Annals of Applied Statistics, vol. 8, no. 1, pp. 176-203, 2014.

[91] L. Xu, "Bayesian Methods for Genetic Association Studies," Toronto, 2012.

[92] M. C. Nikolov, B. A. Coull and P. J. Catalano, "An informative Bayesian structural equation model to assess source-specific health effects of air pollution," Biostatistics, vol. 8, no. 3, pp. 609-624, 07 2007.

[93] E. Stojanovski and K. Mengersen, "Bayesian Structural Equation Models: A Health Application," 2005. [Online]. Available: http://www.mssanz.org.au/modsim05/papers/stojanovski.pdf. [Accessed 11 06 2014].

[94] T. H. H. S. I. I. Sasaki T, "Warfarin-dosing algorithm based on a population pharmacokinetic/pharmacodynamic model combined with Bayesian forecasting.," Pharmacogenomics, vol. 10, no. 8, pp. 1257-66, 2009.

[95] MRC Biostatistics Unit, Cambridge, "Bayesian inference Using Gibbs Sampling Web Site," 2003. [Online]. Available: http://www.mrc-bsu.cam.ac.uk/software/bugs/. [Accessed 11 06 2014].

[96] D. Spiegelhalter, A. Thomas, N. Best and D. Lunn, "WinBUGS User Manual v1.4," MRC Biostatistics Unit, 2003.

[97] Duke University, "Mixture Models and Gibbs Sampling," 2010.

[98] MRC Biostatistics Unit, Cambridge, "WinBUGS Examples," 2003. [Online]. Available: http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-new-

winbugs-examples/. [Accessed 11 06 2014].

[99] Stanford University, "Resources for Learning AMOS," 2010.

[100] B. M. Byrne , Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming, Second Edition, 2nd Edition ed., New York: Routledge, 2010.

[101] IBM Corporation, "IBM SPSS Advanced Statistics 21," 2012.

[102] IBM Corporation, "IBM SPSS Statistics 21 Core System User Guide," 2012.

[103] IBM Corporation, "IBM SPSS Missing Values 21," 2012.

[104] J. Honaker, G. King and M. Blackwell, "Amelia II: A Program for Missing Data," 2012. [Online]. Available: http://gking.harvard.edu/amelia/.

[105] M. Humphries, "Missing Data & How to Deal: An overview of missing data," 2010.

[106] D. C. Howell, "Treatment of Missing Data," University of Vermont, 2012. [Online]. Available: http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html. [Accessed 11 06 2014].

[107] T. D. Pigott, "A Review of Methods for Missing Data," Educational Research and Evaluation: An International Journal on Theory and Practice, vol. 7, no. 4, pp. 353-383, 2001.

[108] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, 2005.

[109] H. Peng, "mRMR (minimum Redundancy Maximum Relevance Feature Selection)," Howard Hughes Medical Institute, Janelia Farm Research Campus, 2007. [Online]. Available: http://penglab.janelia.org/proj/mRMR/.

[110] A. Gelman, "Prior distribution," in Encyclopedia of Environmetrics, vol. 3, Chicheste, John Wiley & Sons, 2002, p. 1634–1637.

[111] D. Kaplan and S. Depaoli, "Bayesian Structural Equation Modeling," in Handbook of Structural Equation Modeling, R. H. Hoyle, Ed., New York, Guilford Publications, 2012, pp. 650-673.

[112] A. J. Carcas and A. M. Borobio, "Efficiency and effectiveness of the use of an acenocoumarol pharmacogenetic dosing algorithm versus usual care in patients with venous thromboembolic disease initiating oral anticoagulation: study protocol for a randomized controlled trial," Trials, vol. 13, no. 239, 2012.

[113] D. F. Wright and S. B. Duffull, "A Bayesian dose-individualization method for warfarin," Clinical Pharmacokinetics, vol. 52, no. 1, pp. 59-68, 2013.

APPENDICES

APPENDIX A: METADATA OF THE IWPC DATA SET

Table 15: Metadata of the IWPC Data Set

Name	Description	Data Type	Unit Of Measure	Unit Of Measure Type
PharmGKB Subject ID	Subject ID numbers in the PharmGKB. A search can be performed to find genotype information on subjects using these ID numbers.	CH[12]		
PharmGKB Sample ID	Sample ID numbers in the PharmGKB. A search can be performed to find genotype information on subjects using these ID numbers.	CH		
Project Site	Coded project site where data was collected.	CH		
Gender	Male, Female or not known = -99	CH		
Race (Reported)	Self-reported information.	CH		
Race (OMB)	Racial categories used are as defined by the Office of Management and Budget, which can be found at http://grants2.nih.gov/grants/guide/notice-files/NOT-OD-01-053.html	CH		
Ethnicity (Reported)	Self-reported information	CH		
Ethnicity (OMB)	Ethnicity categories used are as defined by the Office of Management and Budget, which can be found at http://grants2.nih.gov/grants/guide/notice-files/NOT-OD-01-053.html	CH		
Age	Binned age reported in years (0 - 9, 10 - 19, 20 - 29, 30 - 39, 40 - 49, 50 - 59, 60 - 69, 70 - 79, 80 - 89, 90+)	NR[13]		
Height (cm)	Reported in centimeters	NR	cm	Height
Weight (kg)	Reported in kilograms	NR	kg	Mass
Indication for Warfarin Treatment	DVT = 1, PE = 2, Afib/flutter = 3, Heart Valve = 4, Cardiomyopathy/LV Dilation = 5, Stroke = 6, Post-Orthopedic = 7, Other = 8 or NA; multiple indications are separated by semi-colons	CH		
Comorbidities	List of diseases co-occuring in the patient	CH		
Diabetes	yes = 1, not present = 0 or not known = NA	CH		
Congestive Heart Failure and/or	yes = 1, not present = 0 or not known = NA	CH		

[12] CH stands for Character
[13] NR stands for NUMBER.

Cardiomyopathy				
Valve Replacement	yes = 1, not present = 0 or not known = NA	CH		
Medications	List of medicines taken separated by semi-colons	CH		
Aspirin	yes = 1, not present = 0 or not known = NA	CH		
Acetaminophen or Paracetamol (Tylenol)	yes = 1, not present = 0 or not known = NA	CH		
Was Dose of Acetaminophen or Paracetamol (Tylenol) >1300mg/day	yes = 1, no = 0	CH		
Simvastatin (Zocor)	yes = 1, not present = 0 or not known = NA	CH		
Atorvastatin (Lipitor)	yes = 1, not present = 0 or not known = NA	CH		
Fluvastatin (Lescol)	yes = 1, not present = 0 or not known = NA	CH		
Lovastatin (Mevacor)	yes = 1, not present = 0 or not known = NA	CH		
Pravastatin (Pravachol)	yes = 1, not present = 0 or not known = NA	CH		
Rosuvastatin (Crestor)	yes = 1, not present = 0 or not known = NA	CH		
Cerivastatin (Baycol)	yes = 1, not present = 0 or not known = NA	CH		
Amiodarone (Cordarone)	yes = 1, not present = 0 or not known = NA	CH		
Carbamazepine (Tegretol)	yes = 1, not present = 0 or not known = NA	CH		
Phenytoin (Dilantin)	yes = 1, not present = 0 or not known = NA	CH		
Rifampin or Rifampicin	yes = 1, not present = 0 or not known = NA	CH		
Sulfonamide Antibiotics	Includes Septra, Bactrim, Cotrim and Sulfatrim; yes = 1, not present = 0 or not known = NA	CH		
Macrolide Antibiotics	Includes erythromycin, azithromycin, and clarithromycin; yes = 1, not present = 0 or not known = NA	CH		
Anti-fungal Azoles	Includes ketoconazole, fluconazole, itraconazole, metronidazole, etc. Please do not include other drugs that end in "azole" such as omeprazole or metronidazole; yes = 1, not present = 0 or not known = NA	CH		
Herbal Medications, Vitamins, Supplements	Includes garlic, ginseng, danshen, donquai, vitamins, zinc, iron, magnesium, etc. yes = 1, not present = 0 or not known = NA	CH		
Target INR	Target International Normalized Ratio or NA	CH	mg/week	Mass per Unit Time
Estimated Target INR Range Based	If the target INR is not known, please give estimated target INR ranged based on Indication	NR		

on Indication				
Subject Reached Stable Dose of Warfarin	yes = 1, no = 0 or not known = NA	CH		
Therapeutic Dose of Warfarin	Dose given in milligrams/week	NR	mg/week	Mass per Unit Time
INR on Reported Therapeutic Dose of Warfarin	International Normalized Ratio on the Therapeutic Dose of Warfarin Reported Above	NR		
Current Smoker	yes = 1, not present = 0 or not known = NA	CH		
Cyp2C9 genotypes	*1, *2, *3, *4, *5, *6, *7, *8, *9, *10, *11, *12, or *13 (see https://www.pharmgkb.org/do/serve?objId=PA126&objCls=Gene for specifics of named alleles)	CH		
Genotyped QC Cyp2C9*2	*1, *2, (see https://www.pharmgkb.org/do/serve?objId=PA126&objCls=Gene for specifics of named alleles)	CH		
Genotyped QC Cyp2C9*3	*1, *3 (see https://www.pharmgkb.org/do/serve?objId=PA126&objCls=Gene for specifics of named alleles)	CH		
Combined QC CYP2C9	Combined separate genotypes for *2 and *3 into single CYP2C9 diplotypes with possible values *1/*1, *1/*2, *1/*3, *2/*2, *2/*3, *3/*3	CH		
VKORC1 genotype: -1639 G>A (3673); chr16:31015190; rs9923231; C/T	A/A, A/G, G/G or NA	CH		
VKORC1 QC genotype: -1639 G>A (3673); chr16:31015190; rs9923231; C/T	A/A, A/G, G/G or NA	CH		
VKORC1 genotype: 497T>G (5808); chr16:31013055; rs2884737; A/C	G/G, G/T, T/T or NA	CH		
VKORC1 QC genotype: 497T>G (5808); chr16:31013055; rs2884737;	G/G, G/T, T/T or NA	CH		

A/C				
VKORC1 genotype: 1173 C>T(6484); chr16:31012379; rs9934438; A/G	C/C, C/T, T/T or NA	CH		
VKORC1 QC genotype: 1173 C>T(6484); chr16:31012379; rs9934438; A/G	C/C, C/T, T/T or NA	CH		
VKORC1 genotype: 1542G>C (6853); chr16:31012010; rs8050894; C/G	C/C, C/G, G/G or NA (Note: alleles for 1542 G/C defined on coding strand, while rs8050894 alleles defined on non-coding strand)	CH		
VKORC1 QC genotype: 1542G>C (6853); chr16:31012010; rs8050894; C/G	C/C, C/G, G/G or NA (Note: alleles for 1542 G/C defined on coding strand, while rs8050894 alleles defined on non-coding strand)	CH		
VKORC1 genotype: 3730 G>A (9041); chr16:31009822; rs7294; A/G	A/A, A/G, G/G or NA	CH		
VKORC1 QC genotype: 3730 G>A (9041); chr16:31009822; rs7294; A/G	A/A, A/G, G/G or NA	CH		
VKORC1 genotype: 2255C>T (7566); chr16:31011297; rs2359612; A/G	C/C, C/T, T/T or NA	CH		
VKORC1 QC genotype: 2255C>T (7566); chr16:31011297; rs2359612; A/G	C/C, C/T, T/T or NA	CH		
VKORC1 genotype: -4451 C>A (861); Chr16:3101800	A/A, A/C, C/C or NA	CH		

2; rs17880887; A/C				
VKORC1 QC genotype: -4451 C>A (861); Chr16:31018002; rs17880887; A/C	A/A, A/C, C/C or NA	CH		
CYP2C9 consensus	Derived consensus between original and Combined QC genotypes as follows: If Original equals QC, use common value If Original equals NA, use QC If QC equals NA, use Original If Original is not equal to QC, set value to NA unless Original contains an allele not typed in QC, then use Original All subjects not included in QC genotyping retain Original value	CH		
VKORC1 -1639 consensus	Derived consensus between original and QC genotypes as follows: If Originalequals QC, use common value If Original equals NA, use QC If QC equals NA, use Original If Original is not equal to QC, set value to NA All subjects not included in QC genotyping retain Original value	CH		
VKORC1 497 consensus	Derived consensus between original and QC genotypes as follows: If Original equals QC, use common value If Original equals NA, use QC If QC equals NA, use Original If Original is not equal to QC, set value to NA All subjects not included in QC genotyping retain Original value	CH		
VKORC1 1173 consensus	Derived consensus between original and QC genotypes as follows: If Original equals QC, use common value If Original equals NA, use QC If QC equals NA, use Original If Original is not equal to QC, set value to NA All subjects not included in QC genotyping retain Original value	CH		
VKORC1 1542 consensus	Derived consensus between original and QC genotypes as follows: If Original equals QC, use common value If Original equals NA, use QC If QC equals NA, use Original If Original is not equal to QC, set value to NA All subjects not included in QC genotyping retain Original value	CH		

VKORC1 3730 consensus	Derived consensus between original and QC genotypes as follows: If Original equals QC, use common value If Original equals NA, use QC If QC equals NA, use Original If Original is not equal to QC, set value to NA All subjects not included in QC genotyping retain Original value	CH		
VKORC1 2255 consensus	Derived consensus between original and QC genotypes as follows: If Original equals QC, use common value If Original equals NA, use QC If QC equals NA, use Original If Original is not equal to QC, set value to NA All subjects not included in QC genotyping retain Original value	CH		
VKORC1 -4451 consensus	Derived consensus between original and QC genotypes as follows: If Original equals QC, use common value If Original equals NA, use QC If QC equals NA, use Original If Original is not equal to QC, set value to NA All subjects not included in QC genotyping retain Original value	CH		
Comments regarding Project Site Dataset	Any additional comments that the investigators felt should be included in the release of their data	CH		

APPENDIX B: DESCRIPTIVE STATISTICS FOR SELECTED FEATURES FROM IWPC DATA SET

Table 16: Descriptive Statistics for Selected Features from IWPC Data Set

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std Err	Statistic	Std. Err
AGE	5658	1	9	5.92	1.473	-.609	.033	.069	.065
BSA	4523	1.12	3.38	1.8985	.29896	.567	.036	.655	.073
PE	5700	0	1	.08	.275	3.033	.032	7.199	.065
STROKE	5700	0	1	.04	.192	4.804	.032	21.087	.065
CARDIAC INDICATIONS	5700	0	1	.41	.492	.364	.032	-1.868	.065
DVT	5700	0	1	.09	.291	2.797	.032	5.823	.065
INTERVENTIONS AND SURGERY	5700	0	1	.21	.410	1.396	.032	-.051	.065
TARGET_INR	4871	1.30	99.00	2.5590	1.39237	68.262	.035	4728.866	.070
CANCER	278	0	1	.82	.385	-1.676	.146	.815	.291
DIABET	3572	0	1	.17	.380	1.712	.041	.930	.082
CARDIAC FAILURES	4432	0	1	.20	.399	1.514	.037	.291	.074
VALVE REPLACEMENT	4239	0	1	.24	.427	1.220	.038	-.512	.075
RACE	5470	1	4	1.74	.944	.658	.033	-1.273	.066
RIFAMPIN	2423	0	1	.00	.041	24.566	.050	601.996	.099
AMIODARONE	4182	0	1	.07	.249	3.490	.038	10.182	.076
FLUCONAZOLE	2426	0	1	.01	.083	11.827	.050	138.000	.099
FLUVASTATIN	2423	0	1	.00	.070	14.113	.050	197.331	.099
LOVASTATIN	2431	0	1	.01	.116	8.412	.050	68.824	.099
ASPIRIN	3840	0	1	.24	.426	1.227	.040	-.494	.079
SMOKING	3220	0	1	.14	.346	2.083	.043	2.339	.086
VKORC1 -1639G>A	4201	1	3	1.94	.804	.103	.038	-1.447	.076
VKORC1 1173C>T	3591	1	3	1.85	.824	.290	.041	-1.469	.082
Therapeutic Dose of Warfarin	4837	2.50	315.00	31.4621	16.82898	2.493	.035	23.067	.070
INR on Therapeutic Dose	4968	.80	6.10	2.3644	.46465	.000	.035	1.975	.069
CYP2C9*3	5696	1	99	3.70	15.679	5.913	.032	32.990	.065
CYP2C9*2	5699	1	99	3.75	15.667	5.912	.032	32.995	.065
Valid N (listwise)	87								

APPENDIX C: THE FEATURES OF THE TURKISH DATA SET

Table 17: The Features of Turkish Data Set

Feature Name	Description	Type	Missing	Used
Age	Derived From Year of Birth	CAT[14]l	None	Yes
BSA	Derived From Weight and Height	NUM[15]	None	Yes
Gender	Male/Female	BIN[16]	None	No
Indications	Categorized into following: X4(PE) X5(STROKE) X6(CARDIAC FAILURES) X7(DVT) X8(INTERVENTIONS & SURGERY)	BIN	None	Yes
Medications	Amiodarone and Aspirin are coded as different binary features	-	None	Yes
SMOKER	-	BIN	None	Yes
CYP2C9*2		CAT	None	Yes
CYP2C9*3		CAT	None	Yes
VKORC1:-1639 G		CAT	None	Yes
VKORC1:-1173 C		CAT	None	Yes
CYP4F2		CAT	None	Yes
EPHX1		CAT	None	No
VITAMIN K INTAKE		CAT	None	No
TEA		CAT	None	No
COFFEEA		CAT	None	No
GRAPEFRUIT		CAT	None	No
ALCOHOL		CAT	None	No

[14] Stands for CATEGORICAL.
[15] Stands for NUMBER.
[16] Stands for BINARY.

APPENDIX D: LESSONS LEARNT FOR BAYESIAN ESTIMATION USING AMOS

1. Data processing should be carried out with up-most care. At the end of the day Bayesian SEM is a data driven method.

2. Inspection of the feature set is very crucial. Features can be eliminated by looking at the descriptive statistics and subjective evaluation based on the previous studies.

3. The approach for Missing Value Processing should be determined and justified carefully.

4. Multiple Imputation is a very useful technique. But, the results should be evaluated carefully.

5. A further feature reduction technique can be applied before moving to the phase of Bayesian Inference.

6. Dichotomous (binary) and categorical variables should be processed carefully. Categorical and binary variables do not generally conform to normality principle. Discrete data can be treated as manifestations of an underlying normal distribution. SPSS and AMOS provide a recoding method.

7. Bayesian SEM works for small sample sizes but informative priors should be used and assessed carefully. Prior Sensitivity is higher in small samples.

8. Various software packages are available. WinBUGS does not provide an enhanced user interface for developing the model graphically. It requires programming skills and the documentation is restricted.

9. AMOS is easier to use, has a shorter learning curve and is integrated with SPSS.

10. Dichotomous (binary) and categorical variables should be processed carefully. Categorical and binary variables do not generally conform to normality principle. Discrete data can be treated as manifestations of an underlying normal distribution. SPSS and AMOS provide a recoding facility.

11. Modeling is a hard task. Advance incrementally starting with continuous variables. If all the variables are incorporated at once, the model has the risk of not being identified.

12. Endogenous variables require error terms. These error terms have unstandardized regression weight of 1. Do not forget to assign the weights.

13. Once the model is identified and Bayesian estimation is performed, assess for convergence and the results of standard errors, p values and t-tests etc.

14. A composite structure is obtained by a latent variable with zero variance. For

these latent variables, the regression weight of one of the incoming paths should be assigned to 1 for model identification.

15. Test of non-linearity is applied before the construction of the composite.

16. Always beneficial to use the output «Notes for Model».

17. Chi-square is the most commonly used measure for absolute fit. Always pay attention to it for model evaluation. RMSEA and BIC are other commonly used measures for model fit.

18. For comparing two non-nested models, the BCC (Browne-Cudeck Criterion) can be used for model selection.

19. Selecting Admissibility test sets the prior density to 0 for parameter values that result in a model where any covariance matrix fails to be positive definite.

CURRICULUM VIATE

PERSONAL INFORMATION

Surname, Name: Öztaner, Serdar Murat
Nationality: Turkish (TC)
Date and Place of Birth: 13 October 1968, Ankara
Marital Status: Married
Phone: +90 532 570 40 31
Fax: +90 312 507 79 16
email: murat.oztaner@tcmb.gov.tr

EDUCATION

Degree	Institution	Year of Graduation
Ph.D.	METU Informatics Institute	2014
MS	METU Computer Engineering	1996
BS	METU Computer Engineering	1991
High School	TED Ankara Koleji High School	1987

WORK EXPERIENCE

Year	Place	Enrollment
1998-	Central Bank of the Republic of Turkey	IT Specialist
1996-1998	Netaş	Software Engineer
1991-1996	METU Department of Computer Engineering	Research Assistant

FOREIGN LANGUAGES

Advanced English, Fluent German