Theses and Dissertations

May 2015

Facilitating and Enhancing Biomedical Knowledge Translation: An in Silico Approach to Patient-centered Pharmacogenomic Outcomes Research

Kourosh Ravvaz
University of Wisconsin-Milwaukee

FACILITATING AND ENHANCING

BIOMEDICAL KNOWLEDGE TRANSLATION:

AN *IN SILICO* APPROACH TO PATIENT-CENTERED

PHARMACOGENOMIC OUTCOMES RESEARCH

by

Kourosh Ravvaz

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Biomedical and Health Informatics

at

The University of Wisconsin-Milwaukee

May 2015

ABSTRACT
FACILITATING AND ENHANCING BIOMEDICAL KNOWLEDGE
TRANSLATION: AN *IN SILICO* APPROACH TO PATIENT-CENTERED
PHARMACOGENOMIC OUTCOMES RESEARCH

by

Kourosh Ravvaz

The University of Wisconsin-Milwaukee, 2015
Under the Supervision of Professor Peter J Tonellato

Current research paradigms such as traditional randomized control trials
mostly rely on relatively narrow efficacy data which results in high internal
validity and low external validity. Given this fact and the need to address
many complex real-world healthcare questions in short periods of time,
alternative research designs and approaches should be considered in
translational research. *In silico* modeling studies, along with longitudinal
observational studies, are considered as appropriate feasible means to
address the slow pace of translational research. Taking into consideration this
fact, there is a need for an approach that tests newly discovered gene
variants, via an *in silico* enhanced translational research model (*iS*-TR) to
conduct patient-centered outcomes research and comparative effectiveness
research studies (PCOR CER).

In this dissertation, it was hypothesized that retrospective EMR analysis and
subsequent mathematical modeling and simulation prediction could facilitate
and accelerate the process of generating and translating pharmacogenomic
knowledge on comparative effectiveness of anticoagulation treatment plan(s)
tailored to well defined target populations which eventually results in a
decrease in overall adverse risk and improve individual and population

outcomes. To test this hypothesis, a simulation modeling framework (*iS*-TR) was proposed which takes advantage of the value of longitudinal electronic medical records (EMRs) to provide an effective approach to translate pharmacogenomic anticoagulation knowledge and conduct PCOR CER studies.

The accuracy of the model was demonstrated by reproducing the outcomes of two major randomized clinical trials for individualizing warfarin dosing. A substantial, hospital healthcare use case that demonstrates the value of *iS*-TR when addressing real world anticoagulation PCOR CER challenges was also presented.

iv

DEDICATION AND ACKNOWLEDGMENTS

I view my accomplishments with great humility and thankfulness. Much effort, dedication and sacrifice go into pursuing and earning a PhD, but much more is received. I completed my PhD studies with encouragement, assistance, and support from many wonderful mentors, friends, and family members. I thank my advisor and mentor, Dr. Peter Tonellato for his continuous support, supervision, guidance, knowledge, wisdom, and thoughtful suggestions. I also extend my sincere gratitude to Dr. Timothy Patrick for his mentorship, deep understanding, support, friendship, and great advice throughout the years. I am also grateful to the other members of my committee; Drs. Norma Lang, Huimin Zhao, Michael Michalkiewicz and Chih-Lin Chi as well as other academic mentors at University of Wisconsin-Milwaukee (UWM) for their guidance and support throughout my doctoral program.

A special thanks goes to my friends and colleagues at UWM Laboratory for Public Health Informatics and Genomics and Harvard Medical School Laboratory for Personalized Medicine, present and past, for their fellowship, generous support, intellectual discussions, sharing their valuable experience, and their help throughout the program and the many stages of my dissertation: Chun-Yuan Huang, John Weissert, Charles Murphy, Aurash Mohaimani, John Williams, Vincent Fusaro, Marcus Walz, Emmanuel Owusu Asante-Asamani, Yassine Souilmi, Erik Gafni, Peter Kos, and Prasad Patil. I am equally thankful to my friends and colleagues Carmelo Gaudioso and Lillian Folk for their unwavering support throughout the duration of my time as a graduate student at UWM.

I extend my thanks to Aurora Health Care, Wisconsin for their effective

professional collaboration giving me the great opportunity to conduct my research. I am also grateful to the administrative faculty and staff of the Biomedical and Health Informatics PhD Program who have made the various tasks associated with the doctoral program as smooth as possible.

Finally, I dedicate this dissertation to my family for their endless support and love. My sincere and special thanks to my beloved parents, Heshmat Kamrani and Asghar Ravvaz, for their love, support, and generous sacrifice throughout my life, to my dear wife Leyla for her love and patience, to our lovely daughter Keemia for bringing joy and inspiration to our life, and to my two wonderful, lovely and ardent sisters Azita and Azadeh for their continued support and encouragement.

TABLE OF CONTENTS

Chapter 4: A Comparative Effectiveness Anticoagulation/Anticlotting Research Database: WiAD

Chapter 5: *In Silico* Translational Research Pharmacogenetic Comparative Effectiveness Studies

LIST OF FIGURES

LIST OF TABLES

xiv

Chapter 1: Introduction

1.1.) Statement of the Problem
1.2.) Translational Research, Patient-Centered Outcomes Research and Comparative Effectiveness
1.3.) Conceptual Framework for an Approach that Tests Newly Discovered Genetic Tests, via an *In Silico* Translational Research Model to Conduct Patient-Centered Outcomes Research and Comparative Effectiveness Studies
1.4.) Objective of this work
1.5.) Organization of the Dissertation
1.6.) References

1.1.) Statement of the Problem

Translational research in genomics aims to move promising genomic applications to clinical and public health practice for population health benefit (Cleeren, 2011). Despite demonstrable health benefits of many new genomic discoveries, there remain large barriers between the explosive growth of healthcare related scientific discovery and dramatic improvements in technology and the implementation of this new knowledge and technology to improve health outcomes. There is great optimism that systematic translational research will address these gaps and remove these barriers while appropriately evolving patient-centered care practice to improve the health of individuals and populations (Glasgow, 2012; Waldman, 2010).

A fifteen year study to evaluate predictors of and time required for the translation of highly promising basic research into clinical applications, showed that only about 5% of the basic science findings were licensed for clinical use and only 1% were extensively used for licensed indications (Contopoulos-Ioannidis, 2003).

Since the completion of the Human Genome Project in 2003 (Collins, 2003), advances in genetic science discoveries have led to mounting expectations in regard to their impact on health care and disease prevention. Translating genetic discoveries into lab tests, improved individual care and ultimately into public health improvements, has emerged as an important, but difficult, objective in biomedical research. It is widely recognized that the current translational process is slow, very expensive and often results in an incomplete transfer of research findings into practice, and consequently failure of comparative effectiveness studies used to translate the findings into

substantial changes in patient care and health disparities (Khoury, 2007). Although pharmacogenomics is one of the first clinical applications of the new genomic era, so far, few human genomic discoveries have led to evidence-based applications for medicine and public health and its implementation in clinical practice still involves significant challenges (Swen, 2007; Burke, 2006).

The delay in systematic use of the rapidly expanding collection of clinically valuable genomic discoveries is created by significant problems in the clinical research enterprise including the lack of clinical and biomedical informatic methods, tools and infrastructure required to facilitate the successful translation of the discoveries to practical clinical use. To date, the primary focus of initiatives is to improve the technology, clinical science collaboration and training, and methodologies supporting rapid discovery and regulatory approval of genetic, genomic, and biological markers, associations, and targets. However, efforts to translate discoveries and processes to generate and evaluate evidence in genomic areas (e.g., pharmacogenomics) require prohibitively expensive clinical trial and clinical study validation that are severely hindered by regulatory, technical and validation barriers not easily conducted using current clinical-research or clinical enterprise environments. As an example of these conundrum, highly-sensitive pharmacogenomic (PGx) tests that detect variant alleles combined with increasing genomic knowledge offer physicians the ability to individualize a patient's drug treatment. If pharmacogenomic treatment is successful, one anticipates a large reduction in adverse drug reactions leading to improved patient care, improved outcomes, reduced treatment periods, and overall lower costs.

One notable case is the estimation of initial and managing the maintenance dose of warfarin, the most commonly prescribed oral anticoagulant for the treatment and prevention of thromboembolic events. Many studies have proven warfarin's effectiveness for the prevention of recurrent stroke, ischemic stroke in patients with atrial fibrillation, thromboembolism in patients with mechanical prosthetic heart valves, and myocardial infarction in patients with coronary artery disease (Reynolds, 2007). Warfarin is also effective for the prevention of pulmonary embolism (PE) and deep venous thromboembolism (DVT) in patients requiring orthopedic surgery and in those with a history significant for venous or arterial thromboembolism. Although warfarin remains the therapy of choice, its narrow therapeutic index creates challenges for proper management of anticoagulation, as maintaining the balance of sufficient dosage to prevent thromboembolism while avoiding overdosing to prevent bleeding events is critical. The correct initial dose of warfarin differs widely between individuals with intra-individual variability contributions from factors including age, gender, race, body size, drug interactions, genetics (i.e., mainly VKORC1 and CYP2C9 genes) and compliance. The challenge of warfarin dosing and promise of pharmacogenomics (PGx) have resulted in tens of dosing algorithms including a large number of PGx-based warfarin dosing algorithms.

To date, PGx-based dosing algorithms have not been adequately tested for their impact on clinical outcomes across large hospital diverse patient populations in prospective, controlled trials as it is extremely expensive and time-consuming to conduct the full array of clinical trials required to test and

identify the correct combination of genotypes, phenotypes, clinical and personal data necessary to accurately model drug response, test treatment options and produce the optimal protocol. For example, in most clinical trials the study population is dominated by a specific racial or sub-population group (e.g., about 95% of CoumaGen-I & II study populations were white) which leads to questions on the effectiveness of the PGx testing for different sub-populations and also failure of comparative effectiveness studies and health disparities. Accordingly, there is insufficient evidence, at this time, to recommend for or against routine CYP2C9 and VKORC1 testing in patients under warfarin. Consequently, the use of PGx testing in clinical practices has remained limited. In short, the translation of pharmacogenomic knowledge to clinical practices is associated with challenges and no practical approach to identify the optimal anticoagulation treatment plan exists for large heterogeneous patient populations that accounts for individual risk factors, drug and protocol options, and achieves minimal risk to adverse reactions.

1.2.) Translational Research, Patient-Centered Outcomes Research and Comparative Effectiveness

As mentioned, translational research is designed to move knowledge gained from the basic sciences to its application into clinical and community settings thus improving healthcare outcomes. This process is usually described in phases of translation (i.e., "T-phases"). Recognizing that there are a number of ways to frame the phases, a 5-phase model of translational research process proceeds in iterative and bidirectional phases, research to identify a problem and the discovery of an opportunity or approach to tackle a health issue (T0), research involves basic genome-based discoveries to develop

promising applications such as tests and drugs (T1), research involves evaluating efficacy of such applications and developing evidence-based recommendations (T2), research includes investigations designed to increase uptake and implementation of evidence-based recommendations into practice and public health programs (T3), and research involves evaluation of the effectiveness and cost-effectiveness of genomic applications in the "real world" and in diverse populations (T4) (Khoury, 2012). The translation process is guided by ongoing and updated knowledge synthesis and translation that applies to all phases of translation (Khoury, 2010).

Focusing on these phases, each of which addresses different issues and requires somewhat different methods, provides greater clarity about what is needed if evidence-based approaches are to be successfully implemented and sustained in real-world settings. One of the main methods used in almost all phases from T2 through T4 is conducting clinical trials. Given the complexity and cost of clinical trials, most funded and published genomic research remains in the early phases of translation (Schully, 2012). Consequently, the evidence base for genomics in practice remains limited. In the light of existing challenges in translating pharmacogenetic knowledge of anticoagulants and given the burden of managing anticoagulation therapy using medications with high and variable adverse event risks across diverse populations, there is a clear need for prospective clinical trials that provide direct evidence of the benefits, disadvantages, and costs associated with the genetic testing in the setting of warfarin dosing as well as patient-centered outcomes research and comparative effectiveness studies (PCOR CER). PCOR CER studies are to assist patients, clinicians, and other stakeholders in

making informed decisions by advancing the quality and relevance of evidence concerning the manner in which anticoagulation therapy can effectively and appropriately be managed through research and evidence synthesis that considers variations in patient subpopulations and the dissemination of research findings with respect to the relative health outcomes, clinical effectiveness, and appropriateness of the medical treatments.

Comparative effectiveness research (CER), as a main practical approach to the PCOR, is defined by the Institute of Medicine as "the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care" (IOM, 2009). In the context of anticoagulation therapy, PCOR CER studies could address questions such as: "Given my personal characteristics, conditions and preferences, what should I expect from different anticoagulation therapy protocols?", "What are my anticoagulation therapy options and what are the potential benefits and harms of those options?", "What can I do to improve the outcomes of my anticoagulation therapy given my health condition?", and "How can clinicians and the care delivery systems they work in help me make the best decisions about my anticoagulation-related health?" (PCORI, 2014).

1.3.) An *In Silico* Translational Research Model for Patient-Centered Outcomes Research and Comparative Effectiveness Studies

Current research paradigms such as traditional randomized control trials

mostly rely on relatively narrow efficacy data which results in high internal validity (i.e., extent to which systematic error, bias, is minimized in clinical trials under optimal conditions) and low external validity (i.e., extent to which results of trials provide a correct basis for generalization to other circumstances) (Glasgow, 2012; Nelson, 2006; Juni, 2001; Kessler, 2011). Given this fact and the need to address many complex real-world healthcare questions in short periods of time, alternative research designs and approaches should be considered in translational research. *In silico* modeling studies, along with longitudinal observational studies, are considered as appropriate feasible means to address the slow pace of translational research (Glasgow, 2012). Taking into consideration this fact, there is a need for an approach that tests newly discovered genetic variants, via an *in silico* enhanced translational research model (*iS*-TR) to conduct patient-centered outcomes research and comparative effectiveness studies (Figure 1.1).
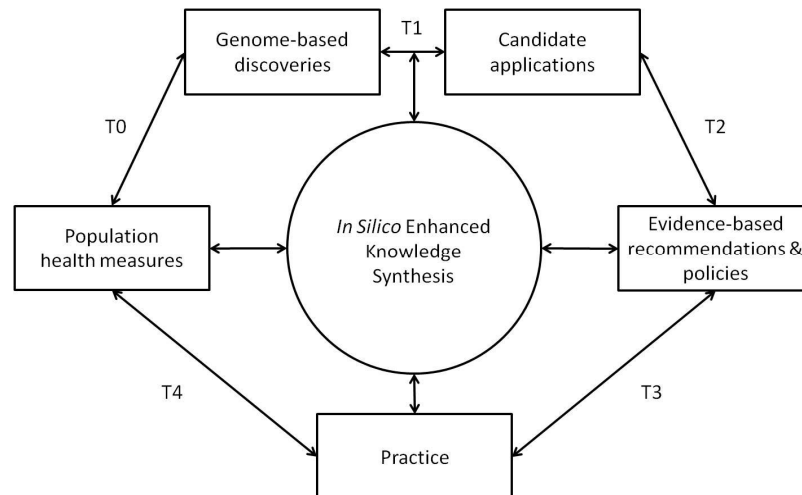
Figure 1.1. Conceptual framework of iS-TR: an iterative and bidirectional translational research framework enhanced with an in Silico knowledge synthesis platform to facilitate pharmacogenetic PCOR and CER studies.

1.4.) Objective of this work:

In this dissertation, we hypothesize that retrospective EMR analysis and subsequent mathematical modeling and simulation prediction can facilitate and accelerate the process of generating and translating pharmacogenomic knowledge on comparative effectiveness of anticoagulation treatment plan(s) tailored to well defined target populations which eventually results in a decrease in overall adverse risk and improve individual and population outcomes. To test this hypothesis, we present a simulation modeling framework, *in silico* enhanced model of translational research (*iS*-TR), which takes advantage of the value of longitudinal electronic medical records (EMRs) to provide an effective approach to translate pharmacogenomic anticoagulation knowledge and conduct PCOR CER studies.

We, first, introduce "*iS*-TR", a translational research model enhanced with *in silico* knowledge synthesis that expedites testing newly discovered genetic variants and eventually facilitates conducting PCOR CER studies (Figure 1.1). Second, we demonstrate the accuracy of the framework by reproducing the

outcomes of two major randomized and clinical effectiveness trials of CoumaGen I and II comparing pharmacogenetic algorithms and standard care for individualizing warfarin dosing. Third, we present a substantial, hospital healthcare Use Case that demonstrates that the value of *iS*-TR when addressing real world anticoagulation PCOR CER challenges.

1.5.) Organization of the Dissertation

This dissertation consists of 7 chapters. An introductory chapter presents the statement of problem and the objectives.

Chapter 2 presents a background on models of translational research, our *in silico* enhanced model of translational research (*iS*-TR), anticoagulation and anticlotting, pharmacogenetics, *in silico* clinical trial studies, and application of the *iS*-TR to a patient-centered pharmacogenetic outcomes research problem,

Chapter 3 describes the details of our *in silico* translational model of genetics testing in patient-centered anticoagulation outcomes research,

Chapter 4 focuses on the details of development of an EMR-based longitudinal comparative effectiveness anticoagulation/anticlotting research database,

Chapter 5 describes a few *in silico* translational research pharmacogenetic comparative effectiveness studies using *iS*-TR,

Chapter 6 presents a PCOR CER study conducted using *iS*-TR and WiAD to address warfarin therapy differences in different subpopulations,

Chapter 7 contains a summary of the important results and conclusions and it also discusses directions for future work in this area using *in silico* translational research.

1.6.) References:

Burke, W., Khoury, M. J., Stewart, A., & Zimmern, R. L. (2006). The path from genome-based research to population health: development of an international public health genomics network. Genetics in medicine : official journal of the American College of Medical Genetics, 8(7), 451-458.

Cleeren, E., Van der Heyden, J., Brand, A., & Van Oyen, H. (2011). Public health in the genomic era: will Public Health Genomics contribute to major changes in the prevention of common diseases? Archives of public health = Archives belges de sante publique, 69(1), 8.

Collins, F. S., Green, E. D., Guttmacher, A. E., & Guyer, M. S. (2003). A vision for the future of genomics research. Nature, 422(6934), 835-847.

Contopoulos-Ioannidis, D. G., Ntzani, E., & Ioannidis, J. P. (2003). Translation of highly promising basic science research into clinical applications. The American journal of medicine, 114(6), 477-484.

Glasgow, R. E., Nelson, C. C., Strycker, L. A., & King, D. K. (2006). Using RE-AIM metrics to evaluate diabetes self-management support interventions. American journal of preventive medicine, 30(1), 67-73.

Glasgow, R. E., Vinson, C., Chambers, D., Khoury, M. J., Kaplan, R. M., & Hunter, C. (2012). National Institutes of Health approaches to dissemination and implementation science: current and future directions. American journal of public health, 102(7), 1274-1281.

Institute of Medicine (IOM). (2009). Initial National Priorities for Comparative Effectiveness Research: The National Academies Press.

Juni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ, 323(7303), 42-46.

Kessler, R., & Glasgow, R. E. (2011). A proposal to speed translation of healthcare research into practice: dramatic change is needed. American journal of preventive medicine, 40(6), 637-644.

Khoury, M. J., Gwinn, M., Dotson, W. D., & Schully, S. D. (2012). Knowledge integration at the center of genomic medicine. Genetics in medicine : official journal of the American College of Medical Genetics, 14(7), 643-647.

Khoury, M. J., Gwinn, M., & Ioannidis, J. P. (2010). The emergence of translational epidemiology: from scientific discovery to population health

impact. American journal of epidemiology, 172(5), 517-524.

Khoury, M. J., Gwinn, M., Yoon, P. W., Dowling, N., Moore, C. A., & Bradley, L. (2007). The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?. Genetics in medicine, 9(10), 665-674.

PCORI. (2014). Patient-Centered Outcomes Research Institute. Retrieved April 20, 2014, from http://www.pcori.org/research-we-support/pcor/

Reynolds KK, V. R., Hartung BR, Linder MW. (2007). Individualizing warfarin therapy. Personalized Medicine, 4(1), 11-31.

Schully, S. D., Benedicto, C. B., & Khoury, M. J. (2012). How can we stimulate translational research in cancer genomics beyond bench to bedside?. Genetics in medicine, 14(1), 169-170.

Swen, J. J., Huizinga, T. W., Gelderblom, H., de Vries, E. G., Assendelft, W. J., Kirchheiner, J., et al. (2007). Translating pharmacogenomics: challenges on the road to the clinic. PLoS medicine, 4(8), e209.

Waldman, S. A., & Terzic, A. (2010). Clinical and translational science: from bench-bedside to global village. Clinical and translational science, 3(5), 254-257.

Chapter 2: Background

2.1.) Translational Research

The progression of scientific knowledge that advances discoveries from "bench to bedside" occurs through a process called translation. The term "translation" which defines and describes the advancement of knowledge through multiple successive phases of research transformation from basic scientific discoveries to public health impact is a complex process that involves different resources and actions and requires both research (e.g., bench-work and clinical trials) and nonresearch activities (e.g., implementation) (Drolet, 2011). The application of findings derived in basic science to the development of new understanding of disease mechanisms, diagnoses, and therapeutics in humans is known as "Translational Research" (Nathan, 2002). Despite an ongoing discussion on the number and nature and stages of translational research, the general consensus is that translational research involves highly iterative and interrelated stages of research in advancing from scientific discoveries to population health (Glasgow, 2012).

In this section, we review the current models and terminology of translation and translation research. We consider the widely adopted models of translational research that have been proposed for different areas of medicine and public health and investigate its applicability to genomic medicine using some examples.

Although the gap between bench and bedside and knowledge translation have been discussed in the last few decades, the translation process has been at the center of attention in biomedical science for only last few years. The Institute of Medicine (IOM) acknowledged the difficulty and importance of translating basic scientific discoveries to clinical applications in its 2001

"Crossing the Quality Chasm" report (Institute of Medicine, 2001). In 2003, the National Institutes of Health (NIH) Roadmap was announced in which translation research was a prominent component (Zerhouni, 2003). Through the Roadmap, many significant programs and major grant efforts have been funded by the NIH to expedite the process and translation. In spite of billions of dollars invested by the NIH to fund basic science research each year, the rate of translation of the results of these research studies into clinical practice has been low and slow. The results of a study on the translation of basic science shows that less than 25% of highly promising biomedical discoveries resulted in at least one published positive randomized clinical trial and less than 5% were established in clinical practice within 20 years (Contopoulos-Ioannidis, 2003). Only 14% of new scientific discoveries entered day-to-day practice and the translation took an average of 17 years (Westfall, 2007). For instance, one study showed that 15 years after successful clinical trials on beta blockers for patients recovering from myocardial infarction, these medications were prescribed for only 62% of patients (Lenfant, 2003). The low percentage of translation, long translational time periods, and low practical implementation would likely be reduced and improved if a known process and clear model of translation of basic science into clinical practice existed and was used. It is vital to identify the continuum of knowledge translation from the laboratory to the point of care. Without having enough understanding of this process, knowledge gets lost in translation and it is difficult to improve the quality of translation and therefore to reach public health gains.

2.1.1.) Models of Translational Research

The IOM Translational Research Model:

In 2003, the US Institute of Medicine (IOM) put forth the principles and a model for translational research. Their translation model ("Clinical Research Continuum", CRC) consists of a two-phase process:

1- From basic science to clinical science and

2- From clinical science to public health impact.

The CRC model is based on different perceptions of basic science researchers and public health agencies on translational research. The first group believes that translational research involves "the transfer of new understandings of disease mechanisms gained in the laboratory into the development of new methods for diagnosis, therapy, and prevention and their first testing in humans" (Sung, 2003). The second group tends to view translational research as "the translation of results from clinical studies into everyday clinical practice and health decision making" (Sung, 2003). IOM refers to the first phase as "T1" translational research (translation of basic
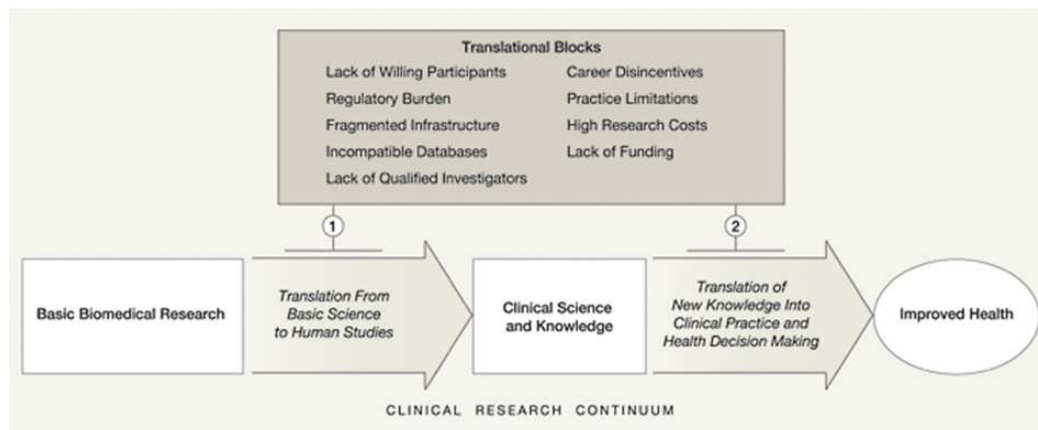
Figure 2.1. The 2 Translational Blocks in the Clinical Research Continuum (Sung, 2003).

research into clinical application) and the second phase as "T2" research (clinical application to evidence-based practice guidelines).

They also identified "Translational Blocks" acting as obstacles in the clinical research continuum in their framework (Figure 2.1).

"Blue Highways" on the NIH Roadmap:

While the standard NIH Roadmap for Medical Research included two major laboratories (bench and bedside) and two above translational steps, Westfall et al., in 2007, divided the second phase into two separate phases resulting in a three-phase translational research model (Figure 2.2) (Westfall, 2007):

1- Basic science to clinical science (T1)

2- Clinical practice (T2)

3- Health improvement (T3)

The following figure displaying this model includes examples of the types of research common in each research laboratory and translational step.
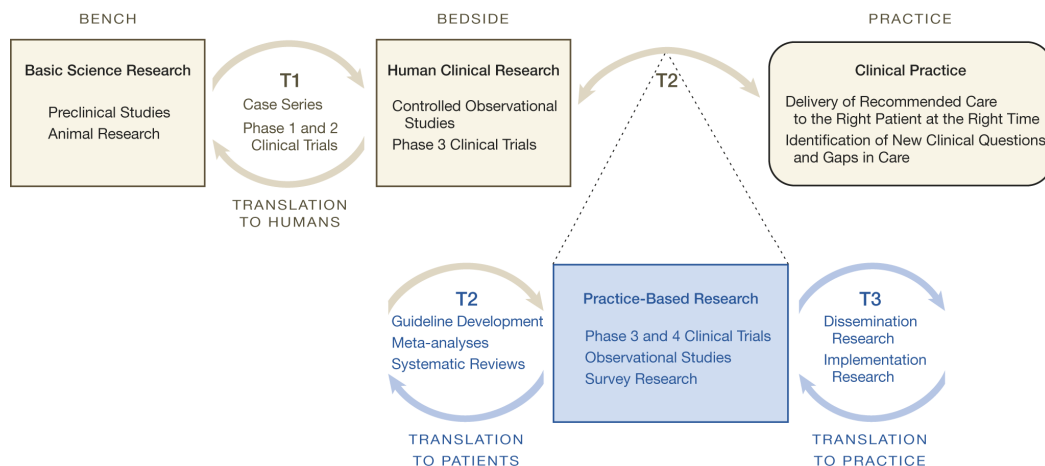
Figure 2.2. Expansion on the NIH Roadmap (Westfall, 2007).

As illustrated in the above figure, the proposed expansion of the NIH Roadmap (blue) consists of (1) an additional research paradigm (Practice-based Research) and (2) a translational step (T3) to improve dissemination

and implementation of research discoveries into clinical practice as the endpoint of the process.

The 3T's Road Map:

Taking into consideration the need for expediting the process of translating and implementing discoveries into clinical settings, a three-phase model of translational research has been offered by Dougherty et al., which addresses the "how" of health care delivery (Figure 2.3) (Dougherty, 2008).
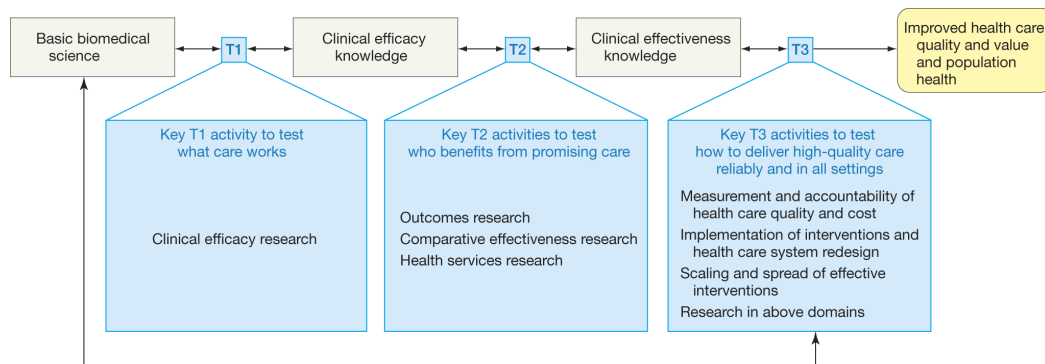
Figure 2.3. The 3T's Road Map - Double-headed arrows represent the essential need for feedback loops between and across the phases of the translational research process (Dougherty, 2008).

This model moves from basic biomedical science to clinical efficacy knowledge (T1). Then, T2 translation focusing on outcomes and comparative effectiveness research results in clinical effectiveness, and patient-centered knowledge and evidence which helps develop individualized treatment plans, more effective "practice guidelines and tools for patients, clinicians and policy makers".

T3 translation activities focus on how new evidence-based treatment, and prevention plans and other interventions are rapidly and reliably incorporated into day-to-day clinical practice and aligned across all levels of the health care system. This phase of translation aimed to improve the health of individuals and populations is accomplished by conducting research in domains such as

measurements, dissemination, and implementation of interventions and healthcare delivery. Policy change is a major component of T3 activities required to enhance health outcomes. In this model, the translational steps T1, T2, and T3 built on each other proceed to improve healthcare delivery over time. This model also includes feedback loops (as represented by the bidirectional arrows in Figure 2.3) to explicitly emphasize on the importance of the bidirectional nature of translational research process.

Biomedical Research Translation Continuum:

Since Drolet et al. believed that none of the prior translation models was unambiguous and the terminology remained indistinct to both researchers and physicians, they proposed a model to define and solidify the concepts and terminology of translation. They called their model "the Biomedical Research Translation Continuum" (Figure 2.4) (Drolet, 2011).

This model has 4 practical landmarks separated by gaps called "Translation Chasms". These chasms represent periods in which translation activities are required to fill the gaps between the phases of research continuum. In the illustration of the model (Figure 2.4), the "zone of translation" depicts the collection of translational tasks that have to be conducted to reach public health impact.

The underlying idea for this model is that a biomedical translation research continuum starts with basic science discoveries that are supposed to be translated to create potential clinical uses. The initial step involves the first translation chasm (T1) in which the basic science discoveries are interpreted in the context of human medical applications. In most cases, addressing T1 chasm requires in vitro laboratory studies and animal models. This is

especially the case for genetic scientific discoveries. So, any new basic discovery entering the continuum is going to be interpreted and translated to human medical applications. The other translation chasms are also bridged similarly. For the T2 chasm, potential human applications are followed by studies such as clinical trials on animal models and humans. In this phase, the safety and efficacy of the interventions based on the new medical applications are evaluated. The output of this phase is proven clinical applications which are going to be implemented and adopted in clinical practice through bridging the T3 chasm. The ultimate goal is to make positive impacts on public health.

To see a complete picture of the translation continuum, they have brought up two examples to examine the entire translation continuum retrospectively. First, it is the example of aspirin for a specific medical application as a medication administered after myocardial infarction (MI) to decrease morbidity and mortality. In this case, initially basic science knowledge from laboratory discoveries (i.e., "acetylsalicylic acid inhibits prostaglandin synthesis") has to be translated to proven clinical practice (administration of aspirin after MI), and, eventually, to individual and public health impacts (decreased mortality). In this process, initially, basic science knowledge has to be translated to a medicine. It happens by bridging the T1 chasm. To cross the first translational chasm (T1), the potential human applications of the medication aspirin should be identified ("Aspirin inhibits platelet aggregation in vivo via inhibition of prostaglandin synthesis") and also biochemical mechanisms that the medication functions in vivo has to be studied in laboratory investigations.

The next step which comes after development of a human application (e.g., "Aspirin prevents post-MI thrombosis by inhibiting platelet aggregation") is to test the safety and efficacy of the medication by conducting clinical research studies such as clinical trials. This translational study crosses the second chasm (T2).

Once the effectiveness of aspirin in decreasing post-MI thrombosis and mortality is demonstrated through clinical trial studies, the medication could be implemented and adopted in clinical practice by bridging the third translational chasm (T3). Then, through practice-based research studies, the public health
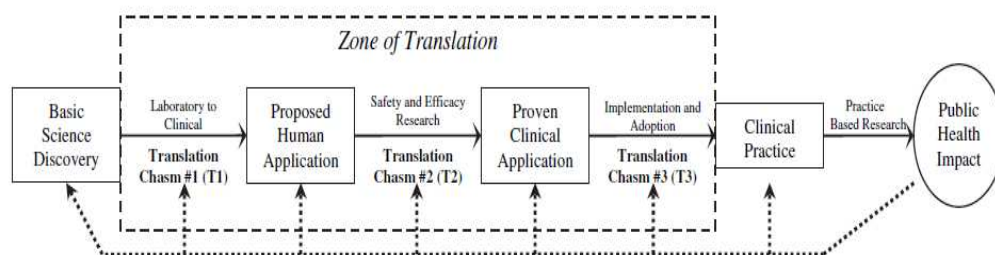
Figure 2.4. Biomedical Research Translation Continuum (Drolet, 2011)

impact of aspirin should be investigated to find out if administering this medication after MI reduces the rate of morbidity and mortality in the population. In this translational model also, the process of translation is bidirectional and "bedside to bench" feedback loops are considered as means to allow integration of new knowledge and also continual improvement of translation process.

The second example is about the administration of beta blockers after MI which is well depicted in the following table (Table 2.1).

Table 2.1. Research Translation Continuum for the administration of beta blockers after MI (Drolet, 2011).

	Pathway	*Inquiry* and Action
Basic Science Discovery	Beta-adrenergic stimulation increases cardiac automaticity and conduction velocity	*Does beta-blockade decrease chronotropy and ionotropy?*
T1	Translation of basic science to humans	Evaluation of biochemical findings in animals, proposal of a potential medical application
Proposed Human Application	Beta-blockade decreases cardiac work and oxygen demand	*Can beta-blocking drugs be used to prevent further ischemic injury following myocardial infarction?*
T2	Translation to clinical treatment (e.g. drug development)	Evaluation of safety and efficacy (i.e. clinical trials)
Effective Clinical Application	Beta-blockers are cardioprotective; safely decreasing post-MI ischemic injury	*How do we get physicians to use beta-blockers in practice?*
T3	Translation to practice	Implementation and Adoption
Clinical Practice	Beta-blocker administration following myocardial infarction	*In true clinical setting, does beta-blocker administration decrease morbidity and mortality?*
**	Practice Based Research	Practice-based research networks; patient registries; cohort and case-control studies; meta-analysis
Public Health Impact	Beta-blockers decrease post-MI mortality; established standard of care	*Can we improve how any part of our treatment pathway (e.g. improved pharmacology, greater adoption, etc)?*
**	*Continual practice improvement*	Findings in any stage feedback to previous research stages (dotted line) for further examination and action

The Continuum of Translation Research in Genomic Medicine:

In the current omics era, it is believed that recent improvements and advances in human genomics and related fields would lead in the future to (1) accelerating the use of new biomarkers derived from gene expression, proteomic, and other omic technologies, and (2) more genomic applications for personalized medicine and disease prevention. Although, there is currently a high interest in evaluating genetic variants for their association with common chronic diseases, however, the rate of use of genetic tests in clinical practice and clinical research has increased at a slow pace in last few years (Pagon 2006). To expedite moving genomic discoveries into practice and the delivery of population-level health benefit, Khoury et al. have proposed a translational model that classifies genomic translational research into the four following multidisciplinary phases of translation (Khoury, 2007):

1- from gene discovery to candidate health applications (T1),

2- from health applications to evidence-based guidelines (T2),

3- from evidence-based guidelines to health practice (T3),

4- from practice to population health impact (T4).

Their phase 4 (one more phase to the Westfall's model) represents the

population-level evaluation of health outcomes (Figure 2.5).
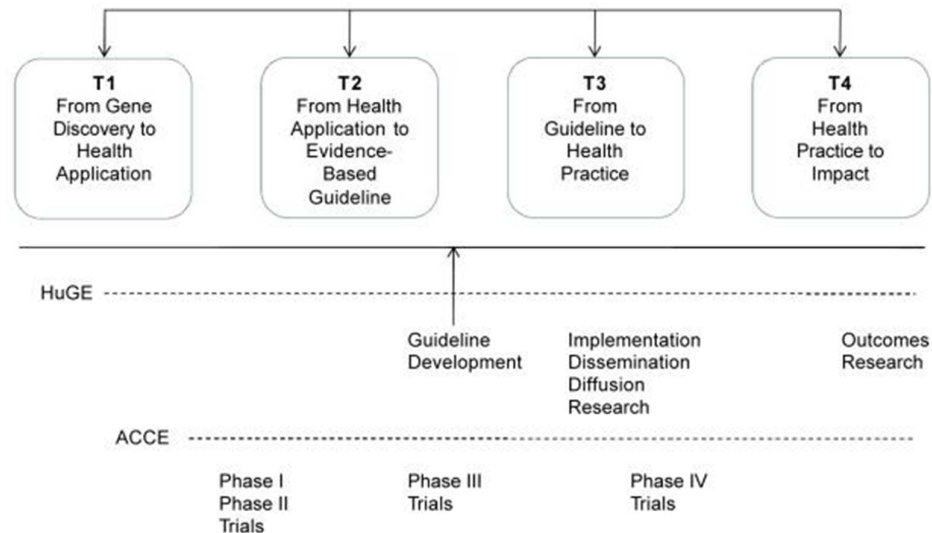
Figure 2.5. The continuum of translation research in genomic
medicine. HuGE, human genome epidemiology; ACCE, analytic
validity, clinical validity, clinical utility, ethical, legal, and social issues.
(Khoury, 2007)

Although the four phase-array of translational research proposed by Khoury et

al. seems to be a linear process, however, this process takes advantage of

feedback loops and it is likely that similar types of research (e.g., clinical trials,

observational studies) are conducted in different phases.

They have also offered the detailed definitions of some terms used in

translational research models (Table 2.2).

Table 2.2. Glossary of certain types of "translation research" involving multiple scientific disciplines (Khoury, 2007).

those that involve observational or clinical trial studies in humans

Human genome epidemiology[26]

Research on prevalence of genetic risk factors, gene-disease associations, and gene-gene and gene-environment interactions in human populations to quantify contribution of genetic factors to human diseases and magnitude of risks

Genetic test evaluation (ACCE components)[27]

A: Analytic validity. Research to measure the ability to accurately and reliably measure the genotype of interest. The four main elements of analytic validity include analytic sensitivity (or the analytic detection rate), analytic specificity (or 1- the analytic false positive rate), laboratory quality control, and assay robustness

C: Clinical validity. Research to measure the test's ability to detect or predict the associated disorder (phenotype)

C: Clinical utility. Research to define the risks and benefits associated with a test's introduction into practice. Specifically, clinical utility focuses on the health outcomes (both positive and negative) associated with testing

E: Ethical, legal, and social issues (some view this component as part of clinical utility). Research to assess concerns specific to genetic information, such as implications for relatives of the person undergoing testing, the possibility of insurance discrimination, and stigmatization based on genotype

Clinical trials[43]

Phase I: Research on a new drug or treatment in a small group of people (20–80) for the first time to evaluate its safety, determine a safe dosage range, and identify side effects

Phase II: The study drug or treatment is given to a larger group of people (100–300) to see whether it is effective and to further evaluate its safety

Phase III: The study drug or treatment is given to large groups of people (1000–3000) to confirm its effectiveness, monitor side effects, compare it to commonly used treatments, and collect information that will allow the drug or treatment to be used safely

Phase IV: The postmarketing studies delineate additional information, including the drug's risks, benefits, and optimal use

Phase 3 translation research (T3)

Dissemination research: Systematic study of how the targeted distribution of information and intervention materials to a specific health audience can be successfully executed so that increased spread of knowledge about the evidence-based interventions achieves greater use and impact of the intervention[24]

Implementation research: Systematic study of how a specific set of activities and designed strategies are used to successfully integrate an evidence-based intervention within specific settings (e.g., primary care clinic, community center, school)[24]

Diffusion research: Systematic study of the factors necessary for successful adoption by stakeholders and the targeted population of an evidence-based intervention that results in widespread use and specifically includes the uptake of new practices or the penetration of broad-scale recommendations through dissemination and implementation efforts, marketing, laws and regulations, systems-research, and policies[24]

Phase 4 translation research (T4)

Outcomes research: Research that describes, interprets, and predicts the impact of various influences, especially (but not exclusively) interventions on "final" endpoints that matter to decision makers. Decision makers may include patients, families, individuals at risk, provider, private and public payers, and so forth[18]

Translational Research and Knowledge Integration:

Glasgow et al., in 2012, argued that although there are "significant advances

in treatments, the public health benefits associated with these improved

treatments tend to be modest because they are not widely implemented"
(Glasgow, 2012). Accordingly, they believe there should be more focus on
some aspects of dissemination and implementation research (rigor and
relevance, efficiency, collaboration, improved capacity, and cumulative
knowledge) to be able to accelerate and improve the appropriate integration
of basic science and genomic discoveries into health care and disease
prevention. Consequently, this prohibits incomplete classification of translation
which might result in missing influential and in essence different tasks that
have to be addressed in the later stages of translation, "dissemination" and
"implementation".

To address this need, they have proposed a different translational research
model with a more differentiated approach to the science of dissemination and
implementation (Figure 2.6). Rather than a linear process of translating
research findings into practice, they have framed the phases of translational
research as a 5-phase model in which research moves into practice and
policy. Having known that research in essence is not a one-way process, they
also believe findings at any phase can impact the other phases and the
translational process is a highly iterative cycle. These 5 overlapping,
interrelated phases span a diverse array of research disciplines, methods and
activities needed to move basic science discoveries to population health.
Each phase moves to progressively broader settings over time and addresses
different questions and requires somewhat different methods to successfully
implement and sustain evidence-based approaches in real-world settings.
Basically, this 5-phase model shows the complexity of translational research

and its inter- and transdisciplinary nature which requires collaboration among partners from basic, clinical, population sciences as well as communities. In the T0 phase, unaddressed health issues and opportunities and the potential scientific approaches to tackle them are identified. Basically, the scientific approaches and discoveries could be derived from multiple disciplines such as molecular, biological, genomic, behavioral, and epidemiological research studies. Then, it is followed by the T1 phase in which promising interventions (e.g., clinical tests, drugs, behavioral, and organizational interventions or policy changes) are tested. In the next phase (T2), the focus is on finding the effectiveness of the new developed interventions and whether they positively impact health outcomes. Different study designs (e.g., clinical trials) and analysis methods are used during this second research phase to create evidence-based recommendations, policies, and guidelines published by respective professional associations and groups. As mentioned in the earlier sections, NIH's primary focus has been on T1 and T2 research and it usually categorizes the activities during T3 and T4 translational research phases under T2. In this model, the T3 phase includes research studies designed and conducted to scale up the implementation of evidence-based recommendations and guidelines into clinical practice settings. And finally, the T4 phase involves real-world evaluation of the population health outcomes of the interventions through different translational studies such as comparative effectiveness and cost effectiveness research studies. Glasgow et al. mainly focus on dissemination and implementation and highly believe in the importance of investigation and understanding of the processes involved in the adoption, implementation, and sustainability of

research. According to them, these translational processes and activities include "dissemination of interventions", "decisions by healthcare practitioners or organizations to adopt or use the interventions", "implementation of the interventions into standard practice or standard operating procedures of organizations", and "maintenance of changes in health care practices by organizations, individual health care practitioners, and patients". In their proposed cyclic model, the translational research process is fueled and
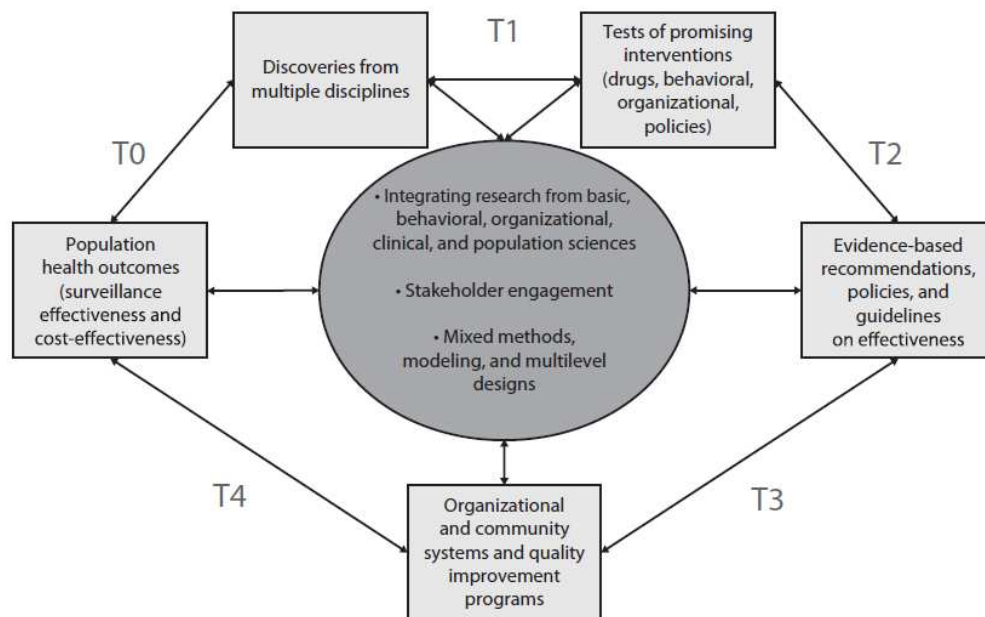
Figure 2.6. 5-phase model of translational research guided by ongoing and updated knowledge synthesis and integration (Glasgow, 2012).

directed by continuing and updated evidence and knowledge synthesis to guide dissemination and implementation research.

2.2.) An *In Silico* Enhanced Model of Translational Research (*iS*-TR)

Given the limitations of the other translational research models and the need to address many complex real-world healthcare questions in short period of

time, alternative research designs, approaches and models should be considered in translational research. *In silico* modeling studies, along with longitudinal observational studies, are considered as appropriate feasible means to address the slow pace of translational research (Glasgow, 2012). Taking into consideration this fact in the context of genomics and specifically pharmacogenomics, there is a need for an approach that tests newly discovered genetic variants, via an *in silico* enhanced translational research model (*iS*-TR) to conduct patient-centered outcomes research and comparative effectiveness studies (PCOR CER). Figure 2.7 depicts the conceptual framework of our proposed *iS*-TR which includes an iterative and bidirectional translational research framework enhanced with an *in silico* knowledge synthesis platform to facilitate pharmacogenetic PCOR and CER studies. Our model and its proposed applications are explained in the following sections.
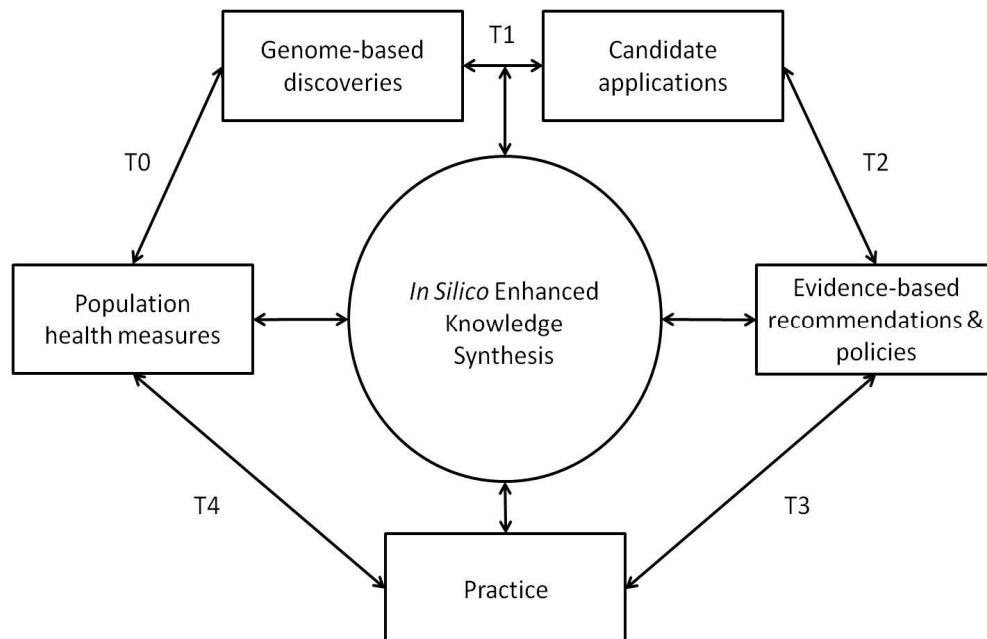
Figure 2.7. Conceptual framework of *iS*-TR

Based on the *iS*-TR model, we have developed an *in silico* translational research framework to facilitate and expedite the pharmacogenomic translational research. This framework consists of a few components including a longitudinal EMR-based anticoagulation research database, a database miner and analyzer, a population health knowledge base, an anticoagulation clinical trial simulator, and a CER Knowledge Base (Figure 2.8). In the following sections, the components of the framework are briefly described.
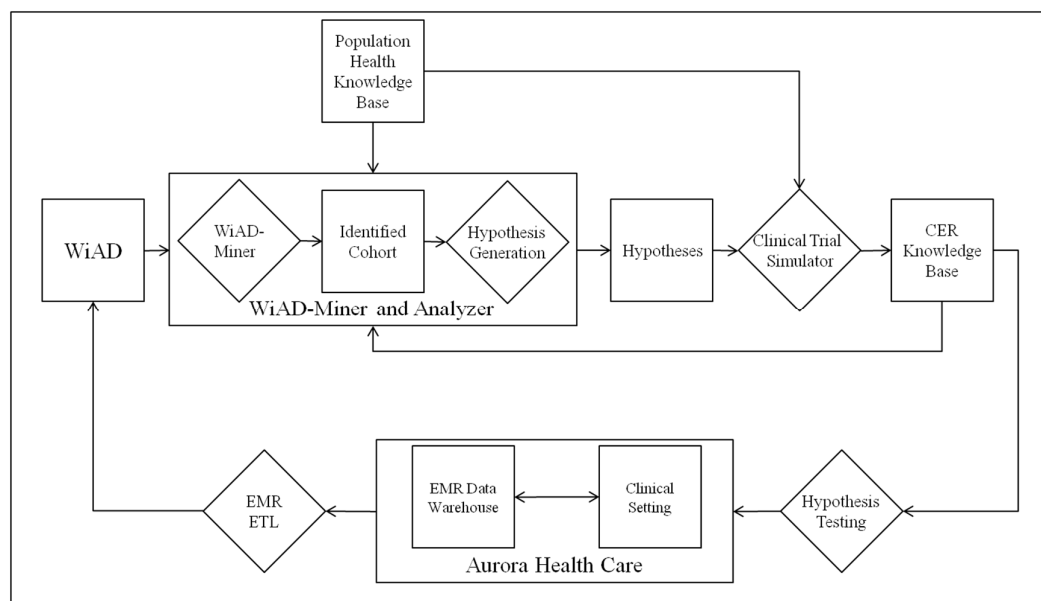
Figure 2.8. *In silico* translational research framework: *In Silico* WiAD PCOR CER Framework. WiAD: Wisconsin Anticoagulation Database; ETL: Extract, Transform, Load.

Wisconsin Anticoagulation Database: WiAD

To generate meaningful PCOR, researchers need high-quality data, including greater clinical detail, longitudinal follow-up, and linkages among data sets (Navathe, 2011). To advance research data infrastructure, University of Wisconsin-Milwaukee and Aurora Health Care Research Institute have collaborated on a multistage project to develop a retrospective EMR-based longitudinal anticoagulation clinical database (Wisconsin Anticoagulation Database: WiAD) being used for PCOR on most frequently prescribed

anticoagulation agents such as Coumadin (Warfarin), Heparin, Ticlopidine (Ticlid), Clopidogrel (Plavix), Dipyridamole (Persantine), Abciximab (ReoPro), Eptifibatide (Integrilin), Tirofiban (Aggrastat), or and Dabigatran (Pradaxa). Aurora Health Care is the largest health care system in Wisconsin operating 15 hospitals throughout the state with more than 3600 licensed beds, 172 physician clinic facilities, and several other health care related entities. It serves about 1.2 million unique patients each year through 7.8 million patient encounters per year. So, such an anticoagulation research database representing Wisconsin State population has provided a powerful tool for conducting outcome research studies on anticoagulation dosing algorithms. The details of this database are described in chapter 4.

WiAD-Miner and Analyzer:

An interactive data profiling and population "segmentation" tool "WiAD-Miner and Analyzer" was developed and used to (a) facilitate the process of patient cohort selection using different demographic, clinical, temporal, and geographical inclusion criteria and (b) synthesize hypotheses. This tool also has some features which are specific to anticoagulation therapy such as a module which calculates the Time in Therapeutic Range (TTR) for each individual from the electronic medical record (EMR) data. WiAD-Miner's details are covered in Chapter 4.

Population Health Knowledge Base:

WiAD is a database that includes data from a geographically widespread, diverse racial and demographic patient population across the state of Wisconsin. A knowledge base including health, demographic and socioeconomic characteristics of original populations of this patient population

provides a rich source of complementary information to be used in different ways for the EMR-based *in silico* PCOR CER studies such as (a) quality control and assurance in the process of transforming and integrating data into the WiAD, (b) generate more accurate virtual patients and patient populations by the clinical trial simulator, (c) generate more enriched hypothesis. The information in the knowledge base is provided from different sources such as the national, state, county and city census data, and state, county and city health reports. This details and usage of this knowledge base is demonstrated in chapters 5 and 6.

Pharmacogenetic Clinical Trial Simulator:

Our pharmacogenetic clinical trial simulator consists of the 5 following adjustable modeling components: 1) A Bayesian network model (BNM) to produce virtual patient population ("Clinical Avatars") consistent with desired target populations, 2) A dose calculator which calculates an initial dose (clinical and PG-based) for each virtual patient, 3) An INR predictor which is based on pharmacokinetic/pharmacodynamics (PK/PD) model, 4) A dose adjuster which adjusts doses by using different protocols based on INRs, and 5) An outcome calculator which measures the desired outcomes (e.g., TTR). The details of this simulator is described in detail in chapter 5.

CER Knowledge Base:

The outcomes of the simulations using the simulator will contribute to the CER knowledge base in comparing effectiveness of different anticoagulation therapy treatment plans and practices from which evidence-based information can be derived by patients, providers, policymakers, and other stakeholders. This knowledge is used as a basis for designing and testing different

population-based treatment plans in real clinical settings. This details and usage of this knowledge base is demonstrated in chapters 5 and 6.

2.3.) Anticoagulation and Anticlotting

The phenomenon of coagulation (thrombogenesis) is a crucial component of the body's hemostasis. Through this process, blood creates clots. Coagulation disorders can result in different forms of bleeding (hemorrhage) and obstructive clotting (thrombosis). Blood coagulation (clotting) is a complex process involving many clotting factors which activates each other. The details of the process are depicted and explained in Figure 2.9.

In brief, this process composes of the three following stages:

- Formation of Prothrombinase

Prothrombinase can be formed either through "intrinsic system" or "extrinsic system" which involves interactions between coagulation factors (e.g., Factor VIII, Factor IX).

- Conversion of Prothrombin to Enzyme Thrombin

In this stage, prothrombinase converts prothrombin to enzyme thrombin.

- Conversion of Fibrinogen to Fibrin (formation of clot)

Then, thrombin converts fibrinogen into fibrin which forms a mesh to form clots.

Various substances are required for the proper functioning of the coagulation cascade (Wikipedia Coagulation, 2014):

- "Calcium and phospholipid (a platelet membrane constituent) are required for the tenase and prothrombinase complexes to function. Calcium mediates
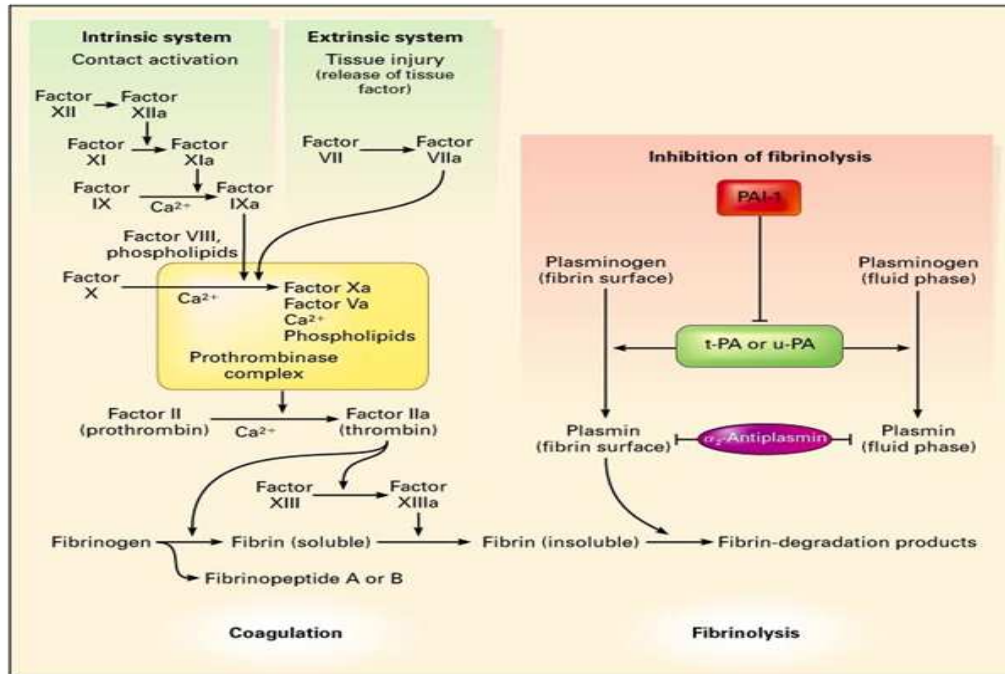
Figure 2.9. The Coagulation and Fibrinolytic Pathways. The main coagulation reactions are divided into the intrinsic and extrinsic systems. Activation of factor XII on contact with a negatively charged surface initiates the intrinsic coagulation system. (The activated form of the factor is indicated by "a.") The extrinsic coagulation system induces the formation of a complex composed of factor VII and tissue factor, which is released after tissue injury. Some of these reactions depend on calcium ions. Thrombin is formed by an enzyme complex called prothrombinase, composed of factor X, factor V, negatively charged phospholipids, and calcium ions. Intrinsic and extrinsic activation of the coagulation cascade leads to the generation of thrombin, the activation of fibrinogen, the release of fibrinopeptides, the formation of soluble fibrin, and finally, the formation of factor XIII–mediated, cross-linked, insoluble fibrin. The main fibrinolytic reactions involve the inhibition of fibrinolysis by plasminogen-activator inhibitor type 1 (PAI-1) and a2-antiplasmin. Fibrinolysis is initiated by tissue plasminogen activator (t-PA), urinary-type plasminogen activator (u-PA), and plasmin. Plasmin bound to the surface of fibrin initiates the lysis of insoluble, cross-linked fibrin, with the subsequent generation of fibrin-degradation products. Plasmin bound to the surface of fibrin is better protected from inhibition by a2-antiplasmin than is plasmin generated in the fluid phase (Kohler; 2000).

the binding of the complexes via the terminal gamma-carboxy residues on

FXa and FIXa to the phospholipid surfaces expressed by platelets, as well as

procoagulant microparticles or microvesicles shed from them. Calcium is also

required at other points in the coagulation cascade.

- Vitamin K is an essential factor to a hepatic gamma-glutamyl carboxylase that adds a carboxyl group to glutamic acid residues on factors II, VII, IX and X, as well as Protein S, Protein C and Protein Z. In adding the gamma-carboxyl group to glutamate residues on the immature clotting factors Vitamin K is itself oxidized. Another enzyme, Vitamin K epoxide reductase, (VKORC) reduces vitamin K back to its active form. Vitamin K epoxide reductase is pharmacologically important as a target of anticoagulant drugs warfarin and related coumarins such as acenocoumarol, phenprocoumon, and dicumarol. These drugs create a deficiency of reduced vitamin K by blocking VKORC, thereby inhibiting maturation of clotting factors. Vitamin K deficiency from other causes (e.g., in malabsorption) or impaired vitamin K metabolism in disease (e.g., in hepatic failure) lead to the formation of PIVKAs (proteins formed in vitamin K absence) which are partially or totally non-gamma carboxylated, affecting the coagulation factors' ability to bind to phospholipid".

Anticoagulation Agents:

Anticoagulation agents are a class of medications that are developed to prevent and reduce blood coagulation and clotting disorders (e.g., DVT: deep vein thrombosis, PE: pulmonary embolism, MI: myocardial infarction and IS:ischemic stroke). Anticoagulation agents are administered in different ways; oral, intravenous, or subcutaneous injection. Different anticoagulants interrupt the coagulation cascade at various points (Figure 2.10). Vitamin K antagonists, such as warfarin, typically work on and inhibit several calcium-dependent clotting factors, including factors II, VII, IX, and X. Dabigatran directly inhibits factor IIa (thrombin). Apixaban, betrixaban, edoxaban, and rivaroxaban inhibit factor Xa.
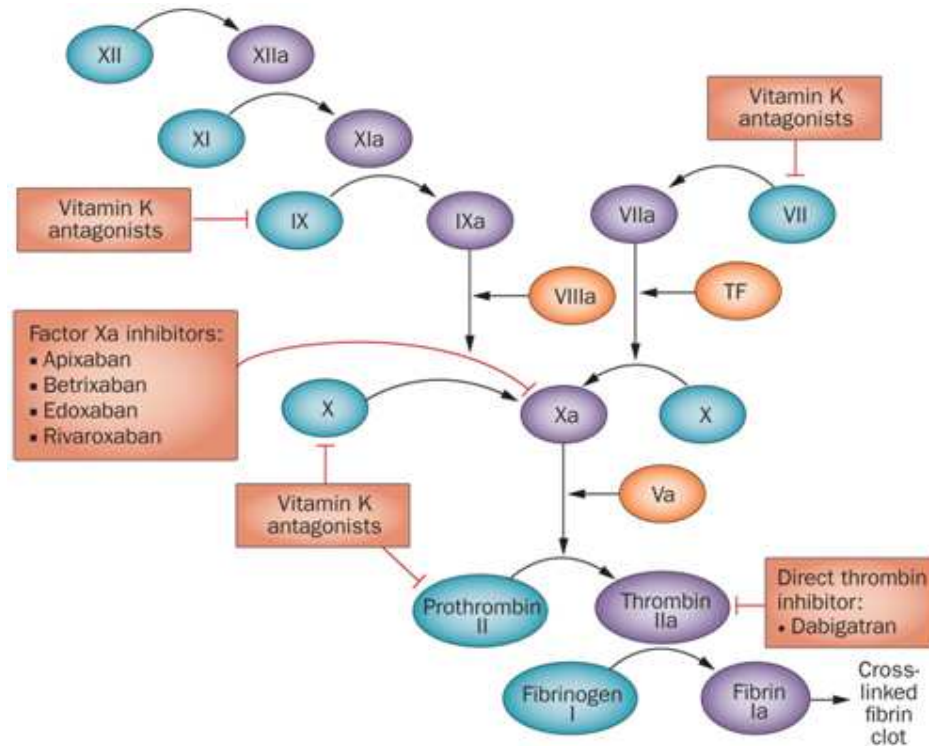
Figure 2.10. Coagulation cascade and point of effect of the common oral anticoagulants. TF: tissue factor. (Makaryus, 2013)

For a few decades, vitamin K antagonists (VKAs) such as warfarin, the most commonly used VKA, have been used as the main agents for long term anticoagulation therapy (Steffel, 2006). Anticoagulation therapy using dose-adjusted VKAs has been always an effective clinical option to prevent and treat thromboembolic diseases. However, long term management of VKAs is challenging as the intensity of anticoagulation represented by measurement of international normalized ration (INR) can be out of desired therapeutic ranges for a large amount of treatment period. Despite the widespread use of VKAs, they have some characteristics that make them difficult to manage, such as (a) a narrow therapeutic index/window outside of which there is a risk of bleeding events, or thromboembolism which demands regular frequent monitoring of INR, (b) a wide inter-individual variability in dose-response due

to their pharmacokinetics affected by genetic and physiological factors, (c) different levels of interactions with other medications and foods (Hart, 2007; Guyatt, 2012).

In addition to VKAs, there are other anticoagulant agents that can be applicable alternatives such as unfractionated heparin, the low molecular-weight heparins (LMWHs), and indirect-acting factor Xa inhibitors (e.g., fondaparinux). However, they have some limitations too. The subcutaneous and parenteral route of administration make them time consuming and less convenient medications for patients under anticoagulation treatment. Furthermore, taking into consideration their specific pharmacokinetics, they need continuous intravenous infusion or daily dose adjustment. Some of them such as LMWHs have unstable bioavailability under some physiological circumstances such obesity or renal failure.

Novel Oral Anticoagulants:

These shortcomings and practical limitations of the VKAs and the intravenous anticoagulant medications have motivated scientists to develop alternative oral medications called novel oral anticoagulants (NOAs) with quick onset of action, predictable pharmacokinetics, less need for regular monitoring and interactions with other medications and foods. NOAs include direct thrombin inhibitors, such as dabigatran, and factor Xa inhibitors such as apixaban, edoxaban, and rivaroxaban (Makaryus, 2013).

The VKAs and NOAs act differently in the body. The first group of medications, inhibit gamma-glutamyl carboxylation of coagulation factors II, VII, IX, X, and the coagulation inhibitor proteins C and S. On the other side, NOAs act on some different proteins in the coagulation cascade (Figure 2.10).

Although NOAs have some benefits over VKAs such as predictable pharmacokinetic mechanisms which facilitate their dosing without a need for routine monitoring, their use have not been popular as expected for some reasons such as their high cost versus VKAs and lack of strategies to rapidly quantify or reverse their anticoagulant effects (Harder, 2008; Brenner, 2011).

2.4.) Pharmacogenetics

Variability in response to medications creates a significant challenge for physicians, patients and pharmaceutical companies (Evans, 1999). Factors involving in the body's response to a medication are multifold and complex (Table 2.3) (Ma, 2011). A large number of clinical studies have shown that variation in genetic make-up of individuals is an important factor affecting the medication response in the body.

Different factors such as environmental effects, physiological factors (e.g., medical conditions) and genetic profile variations are involved in the variation. The field of pharmacogenetics studies the relationship between individual's response variability to medications and genetic variations (Hewett, 2002). This field of study, especially in the light of the complete human genome sequence, has motivated many researchers to conduct pharmacogenetic studies at an accelerating rate in recent years on many medications which were previously recognized to have unpredictable outcomes and unintended side effects.

The knowledge of pharmacogenetics helps to understand some of the underpinning causes of these challenges and also implement personalized medicine. The main questions asked in the field of pharmacogenetics are: what are the genes involved in a drug's mechanism of action? how are a

Table 2.3. Major factors affecting individual medication response (Ma, 2011)

Factors	Effects
Genetic Factors	Major variables; stable and inherited
Therapeutic targets	Drug efficacy (pharmacodynamics)
Drug-metabolizing enzymes	Drug metabolism (pharmacokinetics)
Drug transporters	Drug disposition (pharmacokinetics)
Targets of adverse drug reactions	Drug toxicity (pharmacokinetics and pharmacodynamics)
Factors with indirect effects	Drug efficacy, pharmacokinetics, and toxicity
Other Factors	Mostly transient
Environmental factors	Drug efficacy, pharmacokinetics, and toxicity
Environmental chemicals, coadministered drugs, tobacco smoking, alcohol drinking, and dietary constituents	
Physiological factors	Drug efficacy, pharmacokinetics, and toxicity
Age, sex, disease state, pregnancy, exercise, circadian rhythm, and starvation	

drug's effects propagated through pathways? how can this information be applied to characterize "off-target" adverse events? How can pharmacogenomics information be utilized in prescription and dosing decisions? (Karczewski, 2012).

In clinical settings, physicians mainly prescribe medications based on their clinical judgment and evidence resulting from clinical trials. They usually take into consideration clinical factors (e.g., age, weight, ongoing health condition) and behavioral characteristics of patients and genetic characteristics are not considered in many settings. This appears to be the case in anticoagulation therapy. For example, two groups of patients with similar clinical and backgrounds and presentations might undergo the same dosing regimen of anticoagulant clopidogrel (e.g., 75 mg/day) or warfarin (e.g., 5 mg/day). This

treatment approach might result in a good protection against thromboembolic events in one group and might cause side effects in the other group. Pharmacogenetic studies have revealed that the patients under clopidogrel and warfarin who experience inadequate protection are more likely the ones who are poor metabolizers of the medications owing to their variant alleles of genes CYP2C19 and CYP2C9 and VKORC1, respectively (Highashi, 2002; Aithal, 1999; Rieder, 2005; D'Andrea, 2005; Yuan, 2005; Aquilante 2006). In the case of warfarin, these findings have resulted in a sheer number of pharmacogenetic-based (PG-based) dosing algorithms explaining a significant proportion of the interindividual variability in warfarin dose requirement (e.g., Gage, 2008; Anderson, 2007; Klein 2009).

After taking a medication, it has to go through different components of the body to reach its target tissues/cells, then it acts on its target, and eventually its metabolites and residues are eliminated from the body. The process of absorbing, distributing, metabolizing, and excretion/elimination are regulated by pharmacokinetic (PK) genes. Pharmacodynamic (PD) genes regulate the effect of medications on their targets. Genes regulating PK and PD processes can be involved and led to desired/intended effects by affecting target cells or contribute to undesired/side effects by affecting non-target cells. Pharmacogenetic researchers try to find the genes involved in both the PK and PD pathways that affect drug action in order to improve dosing and avoid adverse drug reactions (Karczewski, 2012).

Different stakeholders such as patients, health care providers, pharmaceutical companies and academics can take advantage and are interested in pharmacogenetic knowledge. Patients and healthcare providers use

pharmacogenetic information to make more informed decisions and determine more accurate medication optimal doses. Using pharmacogenetic discoveries and knowledge, research teams and pharmaceutical companies are able to enhance and facilitate safer and target-oriented clinical trials.

2.5.) *In Silico* Clinical Trial Studies

In silico is a term which is used and referred to tasks performed on computer or through computer simulation. *In silico* techniques and methods have been widely used in different disciplines such as engineering, physics, astronomy, marketing and economics to foster the process of developing and testing the performance of systems. The application of simulation in these fields have resulted in reduction of costs and shorter development cycles.

In last two decades, clinical trials modeling and simulation has gained a lot of attention. *In silico* approaches to conduct clinical trials which employ realistic virtual subjects and typical trial conditions, based on both experimentally informed disease progress and drug intervention models have been embraced by both pharmaceutical companies and also regulatory agencies (Kimko, 2003). It provides an opportunity for researchers to develop and test hypotheses virtually prior to real-world experiments. Many pharmaceutical companies use the clinical trials modeling and simulation techniques to facilitate the development of new drugs and make the drugs more efficient. Regulatory agencies such as US Food and Drug Administration (FDA) and European Medicines Agency have recognized the important role of clinical trials modeling and simulation and have advocated using it to support more evidence-based study designs and dosing protocols in different target subpopulations. FDA underscored clinical trial modeling and simulation in its

2004 Clinical Path Initiative as an opportunity which could improve predictability and efficiency along the critical path from laboratory concept to commercial product: "FDA scientists use, and are collaborating with others in the refinement of, quantitative clinical trial modeling using simulation software to improve trial design and to predict outcomes." (FDA, 2004).

From the perspective of European Medicines Agency, the modeling and simulation of clinical trials and PK/PD data contribute to the regulatory review process and also drug development because modeling and simulation (Jönssen, 2010): "allow more efficient utilization of collected clinical data", "support informed decision making regarding future studies and study designs including dose selection", are beneficial in time and cost savings. Accordingly, several European guidelines "recommend modeling and simulation as a useful tool to support dose selection and establish dose recommendations in special populations".

Peck et al. have provided the following detailed technical definition for the clinical trial simulation: "the generation of biomarker or clinical responses in virtual subjects that take into account (a) the trial design and execution, (b) pathophysiological changes in subjects during the trial (disease progress model), and (c) pharmacology (drug intervention model), using mathematical, statistical and numerical methods and models" (Peck, 2011).

Components of a Clinical Trial Simulation:

A clinical trial simulation generally is composed of three following components (Holford, 2000):

1- The input–output model consists of submodels that incorporate the drug's pharmacokinetics and pharmacodynamics, the disease progression during the

trial, the trial endpoints, and the residual variability. Some of these submodels may include covariate influences on model parameters, which comprise the covariate distribution model. Basically, the input–output models are functions that map the set of inputs to the set of outputs.

2- The covariate distribution model describes the distribution of the covariates and their intercorrelations.

3- The trial execution model consists of the study design elements, and potential submodels for compliance, protocol deviations, and missing data.

2.6.) Application of the *iS*-TR to a Patient-Centered Pharmacogenetic Outcomes Research Problem

Based on the Patient Protection and Affordable Care Act (PPACA, 2010), Comparative Effectiveness Research (CER) is a national research priority. Under this Act, the Patient-Centered Outcomes Research Institute (PCORI) is in charge of supporting the research that takes into account the potential for differences in the effectiveness of health care treatments, services, and items as used with various subpopulations, such as racial and ethnic minorities, women, age, and groups of individuals with different comorbidities, genetic and molecular sub-types, or quality of life preferences and include members of such subpopulations as subjects in the research as feasible and appropriate. Accordingly, PCORI has proposed priorities and research agenda with focus on CER studies providing opportunities to assess the benefits and risks of adopting genetic tests in patient subsets (PCORI, 2012). PCORI has recently published its "Methodology Report" introducing the PCORI Methodology Standards (PCORI, 2013). These are specific recommendations for researchers that designate the minimal requirements for

following PCOR best practices. Under the recommended methods related to Heterogeneity of Treatment Effect (HTE: a technical term for the fact that different people do not always respond the same way to the same treatment), it is recommended to develop methods to use simulation models to (a) address questions on heterogeneity of treatment effect, (b) address patient-centered comparative effectiveness questions, and (c) to support guidance on adaptive trials' complex design specific to PCOR.

Accordingly, we use our *iS*-TR model and framework to generate robust, relevant, and timely evidence for patient-centered pharmacogenetic outcomes research questions. For instance, although the potential clinical value of most of the PG-based algorithms versus non PG-based "best practice" treatment plans was assessed through rigorous randomized controlled trials, their clinical applicability and effectiveness for different target populations have not been evaluated which leads to an opportunity for PCOR CER studies using our *iS*-TR framework.

2.7.) Focus of this study

As part of continuing effort to address health care challenges through patient-centered outcomes research (PCOR), we introduce an *in silico* translational research model and framework supporting pharmacogenetic anticoagulation PCOR CER prediction and validation studies. This framework is designed to demonstrate how current access to large comprehensive electronic medical records (EMR) covering diverse patient populations, coupled with novel modeling and computational simulations could provide an unprecedented opportunity to conduct *in silico* identification, validation and comparison of treatment strategies.

2.8.) References:

Aithal, G. P., Day, C. P., Kesteven, P. J., & Daly, A. K. (1999). Association of polymorphisms in the cytochrome P450 CYP2C9 with warfarin dose requirement and risk of bleeding complications. Lancet, 353(9154), 717-719.

Anderson, J. L., Horne, B. D., Stevens, S. M., Grove, A. S., Barton, S., Nicholas, Z. P., et al. (2007). Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. Circulation, 116(22), 2563-2570.

Aquilante, C. L., Langaee, T. Y., Lopez, L. M., Yarandi, H. N., Tromberg, J. S., Mohuczy, D., et al. (2006). Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. Clinical pharmacology and therapeutics, 79(4), 291-302.

Brenner, B., & Hoffman, R. (2011). Emerging options in the treatment of deep vein thrombosis and pulmonary embolism. Blood reviews, 25(5), 215-221.

Contopoulos-Ioannidis, D. G., Ntzani, E., & Ioannidis, J. P. (2003). Translation of highly promising basic science research into clinical applications. The American journal of medicine, 114(6), 477-484.

D'Andrea, G., D'Ambrosio, R. L., Di Perna, P., Chetta, M., Santacroce, R., Brancaccio, V., et al. (2005). A polymorphism in the VKORC1 gene is associated with an interindividual variability in the dose-anticoagulant effect of warfarin. Blood, 105(2), 645-649.

Dougherty, D., & Conway, P. H. (2008). The "3T's" road map to transform US health care: the "how" of high-quality care. JAMA : the journal of the American Medical Association, 299(19), 2319-2321.

Drolet, B. C., & Lorenzi, N. M. (2011). Translational research: understanding the continuum from bench to bedside. Translational research : the journal of laboratory and clinical medicine, 157(1), 1-5.

Evans, W. E., & Relling, M. V. (1999). Pharmacogenomics: translating functional genomics into rational therapeutics. Science, 286(5439), 487-491.

Food and Drug Administration (FDA). Challenges and Opportunities Report– March 2004. Retrieved August 5, 2014, from http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm

Gage, B., Eby, C., Johnson, J., Deych, E., Rieder, M., Ridker, P., et al. (2008). Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. Clinical Pharmacology & Therapeutics, 84(3), 326-331.

Glasgow, R. E., Vinson, C., Chambers, D., Khoury, M. J., Kaplan, R. M., & Hunter, C. (2012). National Institutes of Health approaches to dissemination and implementation science: current and future directions. American journal of public health, 102(7), 1274-1281.

Guyatt, G. H., Akl, E. A., Crowther, M., Gutterman, D. D., & Schuunemann, H. J. (2012). Executive summary: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest, 141(2 Suppl), 7S-47S.

Harder, S., Parisius, J., & Picard-Willems, B. (2008). Monitoring direct FXa-inhibitors and fondaparinux by Prothrombinase-induced Clotting Time (PiCT): relation to FXa-activity and influence of assay modifications. Thrombosis research, 123(2), 396-403.

Hart, R. G., Pearce, L. A., & Aguilar, M. I. (2007). Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. Annals of internal medicine, 146(12), 857-867.

Hewett, M., Oliver, D. E., Rubin, D. L., Easton, K. L., Stuart, J. M., Altman, R. B., et al. (2002). PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic acids research, 30(1), 163-165.

Higashi, M. K., Veenstra, D. L., Kondo, L. M., Wittkowsky, A. K., Srinouanprachanh, S. L., Farin, F. M., et al. (2002). Association between CYP2C9 genetic variants and anticoagulation-related outcomes during warfarin therapy. JAMA : the journal of the American Medical Association, 287(13), 1690-1698.

Holford, N., Kimko, H., Monteleone, J., & Peck, C. C. (2000). Simulation of clinical trials. Annual review of pharmacology and toxicology, 40(1), 209-234.

Institute of Medicine. (2001). Crossing the quality chasm: A new health system for the 21st century: National Academies Press.

Jönsson, S., Henningsson, A., Edholm, M., & Salmonson, T. (2011). Contribution of modeling and simulation studies in the regulatory review: a European regulatory perspective Clinical Trial Simulations (pp. 15-36): Springer.

Karczewski, K. J., Daneshjou, R., & Altman, R. B. (2012). Chapter 7: Pharmacogenomics. PLoS computational biology, 8(12), e1002817.

Khoury, M. J., Gwinn, M., Yoon, P. W., Dowling, N., Moore, C. A., & Bradley, L. (2007). The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?. Genetics in medicine : official journal of the American College of Medical Genetics, 9(10), 665-674.

Kimko, H., & Duffull, S. B. (2003). Simulation for designing clinical trials: a pharmacokinetic-pharmacodynamic modeling perspective: CRC Press.

Klein, T., Altman, R., Eriksson, N., Gage, B., Kimmel, S., Lee, M., et al. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. New England journal of medicine, 360(8), 753-764.

Kohler, H. P., & Grant, P. J. (2000). Plasminogen-activator inhibitor type 1 and coronary artery disease. The New England journal of medicine, 342(24), 1792-1801.

Lenfant, C. (2003). Shattuck lecture--clinical research to clinical practice--lost in translation? The New England journal of medicine, 349(9), 868-874.

Ma, Q., & Lu, A. Y. (2011). Pharmacogenetics, pharmacogenomics, and individualized medicine. Pharmacological reviews, 63(2), 437-459.

Makaryus, J. N., Halperin, J. L., & Lau, J. F. (2013). Oral anticoagulants in the management of venous thromboembolism. Nature reviews. Cardiology, 10(7), 397-409.

Nathan, D. G. (2002). Careers in translational clinical research-historical perspectives, future challenges. JAMA : the journal of the American Medical Association, 287(18), 2424-2427.

Navathe, A. S., Clancy, C., & Glied, S. (2011). Advancing research data infrastructure for patient-centered outcomes research. JAMA : the journal of the American Medical Association, 306(11), 1254-1255.

Pagon, R. A. (2006). GeneTests: an online genetic information resource for health care providers. Journal of the Medical Library Association: JMLA, 94(3), 343-348.

Patient-Centered Outcomes Research Institute (PCORI). (2012). National priorities for research and research agenda. Retrieved August 5, 2014, from http://www.pcori.org/content/national-priorities-and-research-agenda

Patient-Centered Outcomes Research Institute (PCORI) Methodology Committee. (2013). The PCORI Methodology Report. Retrieved July 7, 2014, from pcori.org/research-we-support/research-methodology-standards.

Peck, C. C., & Kimko, H. H. (2011). Clinical Trial Simulation and Quantitative Pharmacology Clinical Trial Simulations (pp. 1-11): Springer.

Rieder, M. J., Reiner, A. P., Gage, B. F., Nickerson, D. A., Eby, C. S., McLeod, H. L., et al. (2005). Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. The New England journal of medicine, 352(22), 2285-2293.

Steffel, J., Luscher, T. F., & Tanner, F. C. (2006). Tissue factor in cardiovascular diseases: molecular mechanisms and clinical implications. Circulation, 113(5), 722-731.

Sung, N. S., Crowley, W. F., Jr., Genel, M., Salber, P., Sandy, L., Sherwood, L. M., et al. (2003). Central challenges facing the national clinical research enterprise. JAMA : the journal of the American Medical Association, 289(10), 1278-1287.

The Patient Protection and Affordable Care Act of 2010 (PPACA). (2010). 111–1482010. Retrieved August 10, 2014, from http://www.gpo.gov/fdsys/pkg/BILLS-111hr3590enr/pdf/BILLS-111hr3590enr.pdf

Westfall, J. M., Mold, J., & Fagnan, L. (2007). Practice-based research--"Blue Highways" on the NIH roadmap. JAMA : the journal of the American Medical Association, 297(4), 403-406.

Wikipedia contributors. (2014, March 15, 2014). Coagulation. Retrieved April 7, 2014, from http://en.wikipedia.org/w/index.php?title=Coagulation&oldid=599705965

Yuan, H. Y., Chen, J. J., Lee, M. T., Wung, J. C., Chen, Y. F., Charng, M. J., et al. (2005). A novel functional VKORC1 promoter polymorphism is associated with inter-individual and inter-ethnic differences in warfarin sensitivity. Human molecular genetics, 14(13), 1745-1751.

Zerhouni, E. (2003). Medicine. The NIH Roadmap. Science, 302(5642), 63-72.

Chapter 3: Patient-Centered Anticoagulation and Anticlotting Treatment Outcomes Research: An *In Silico* Translational Model of Genetics Testing in Patient Centered Anticoagulation Outcomes Research

3.1.) Patient Populations and Subpopulations

3.1.1.) Clinical Avatar Model

3.2.) Clinical Pharmacology

3.2.1.) Mechanism of Action

3.2.2.) Pharmacokinetics and Pharmacodynamics

3.2.3.) Pharmacogenomics

3.2.4.) Accounting for Adverse Reactions

3.2.5.) Clinical Protocols

3.2.6.) Dosage and Administration

3.2.7.) PK/PD Modeling and Simulation (INR Prediction)

3.3.) Design of Comparative Effectiveness and Clinical Trial Studies

3.3.1.) Longitudinal Studies

3.3.2.) INT and TTR Estimation

3.3.3.) Multiple Protocol Studies

3.4.) Model and Simulation Requirements

3.4.1.) Patient Populations

3.4.2.) Parameters

3.5.) Outcome Metrics

3.5.1.) Primary Outcome Metric

3.5.2.) Secondary Outcome Metrics

3.6.) References

3.1.) Patient Populations and Subpopulations

As discussed in section 1.3, current research paradigms such as traditional randomized control trials mostly rely on relatively narrow efficacy data which results in high internal validity (i.e., extent to which systematic error, bias, is minimized in clinical trials under optimal conditions) and low external validity (i.e., extent to which results of trials provide a correct basis for generalization to other circumstances) (Glasgow, 2006; Juni, 2001; Kessler, 2011). The lack of comparability between trial participants and nonparticipants has resulted in a slow pace of translation of new genomic knowledge to clinical settings and public health. In almost all of the controlled trials, a small fraction of the total number of patients participate. The participants are usually dedicated individuals to trial studies who are selected based on strict inclusion and exclusion criteria. So, the interventions that are successful in trial patient cohorts may not necessarily translate well to the real-world clinical setting. This raises questions of the generalizability of the results of clinical trials which is one of the important and practical aspects of clinical trials (Elting, 2006; Schulz, 2010). There are different approaches that can be taken to answer the questions, "How can trial results be applied to patients in clinical practice?" and "How different are the interventions/treatment plans resulted from a given trial study compared with the other available ones?" For example, (a) conducting large, population-based effectiveness trials to provide evidence on the generalizability of clinical trial results and the realistic benefits of treatments, (b) conducting effectiveness studies that examine how a therapy/intervention, that works effectively under certain circumstances, works in clinical practice, (c) developing trials that fit patients seen in practice and

their participants share the same characteristics as the majority of patients, and (d) another approach is the development of trials that are targeted and appropriate to the needs of special populations. Accordingly, continued investigation is needed to see how results of clinical trials are translated into nontrial care, to facilitate the dissemination of clinical trials findings, and to identify ways to achieve faster and more generalizable clinical trial findings. This has been the case for pharmacogenomics in the last decade too. Enormous number of clinical trials have been conducted or are currently under way to test the accuracy of previous dosing algorithms, construct new dosing algorithms, or test the value of genetic tests in warfarin dosing, with equivocal results. Several clinical trial studies have shown that PG-based dosing lead to superior control of warfarin anticoagulation (e.g., Gage, 2008; Anderson, 2012), whereas a number of prospective studies and controlled clinical trials have failed to show that genotyping improves warfarin dosing and anticoagulation control (e.g., Anderson, 2007; Kimmel, 2013). Although in those clinical trials thousands of subjects have been recruited, the clinical utility of PG-based dosing of warfarin has been mostly tested in small clinical trial or observational study populations.

To deal with such a challenge, overcome the study populations' limitations, and fill the translational research gaps using *in silico* trial and comparative effectiveness studies, in the first step, there is a need to generate realistic virtual subject populations and subpopulations representing the patient types that are likely to be studied in the actual studies. Among various methods, we have used the following methods to create virtual patient populations

(hereafter "Clinical Avatars") in our *in silico* PC-CER translational framework, introduced in section 2.2.

3.1.1.) Clinical Avatar Model

Background in Bayesian Theory

Simulating patient data (i.e., creating "Clinical Avatars") poses several challenges common to big data research including missing values, high dimensionality and low sample size. We found that a Bayesian machine learning approach can address these issues in an efficient, effective and generalizable method. More specifically, our method to generate clinical avatars relies on constructing a Bayesian belief networks, hereafter called Bayesian network model (BNM). Using Bayes theorem and extensions of Bayesian theory, BNMs can accomplish two critical elements necessary for modeling patient data, pattern learning and parameter learning. Pattern learning describes a method for discovering graphical representation that correctly approximates relationships between variables within some set of data. Parameter learning applies a graphical pattern with the names and number of categories within the pattern to create a Bayes net. The Bayes net is composed of distinct states each with unique conditional probabilities as defined by some graphic pattern and estimates provided by either training data or evidence.

As recently summarized by Conca Bielza (2014) Bayesian Classification, (i.e., pattern learning) offers distinct advantages in modeling healthcare data over other statistical classification techniques (e.g., ad hoc, regression). Types of Bayesian classifiers, such as binary, categorical and continuous are used to capture the structure of data sets found in patient databases. Such classifiers

also efficiently accommodate missing data as well as feature selection in the learning and inference stages. Bayesian networks also provide "an explicit, graphical and interpretable representation of uncertain knowledge…based on the sound concept of conditional independence (as) an example of a probabilistic graphical model" (Bielza, 2014). Once the pattern is discovered, the parameters can be learned efficiently. In the following section, the basis and method to create Clinical Avatars are summarized: the fundamental principles of classification, Bayes theorem, Markov Blanket Based DAG Discovery and Bayesian belief networks.

A.) Data Classification

Classification is a broad term used to describe the process of assigning or predicting categorical classes and defining their respective relationships. Classification is a two-step process; the first step is termed the learning step and the second step is termed the classifier step. During the learning step, or training phase, a classification algorithm creates a model by learning from a set of training data (i.e., database tuples) and a set of corresponding classes *C*. A tuple, **X**, is represented by an n-dimensional attribute vector, **X** = (x_1, x_2, …, x_n). **X** is composed of a random set of predictive n-measurements from the tuple of n-database attributes, such that, $a_i \in \Omega\ a_i = \{A_1, A_2, …, A_n\}$. Each tuple, **X**, is assumed to belong to a predefined class as determined by another database attribute called the class label attribute called *C* where $c \in \Omega\ c = \{c_1, c_2, …, c_n\}$. The class label attribute is categorical in that each value of c_n serves as a category or class. *Classification* denotes instances where the class label of each training tuple is provided, are known as supervised learning (i.e., the learning of the classifier is "supervised" in that it is told to

which class each training tuple belongs). In contrast *Clustering* denotes instances where the class label of each training tuple is unknown, known as unsupervised learning and the tuples as organized by a specified measure of similarity. The classification problem consists of inducing a model **M** from a random sample of individual tuples called a training set or training tuples D, where D = {(x$^{(1)}$, c$^{(1)}$),…, {(x$^{(N)}$, c$^{(N)}$)} of size N from a joint probability distribution p(**X**, C).

In the second step, the predictive accuracy of the induced model **M** is estimated. To do so, there are multiple measures such as classification accuracy and sensitivity and ROC curve all of which must validated by bootstrapping, k-fold cross validation, and/or holding-out a validation set of data tuples **X**$_j$ and their associated class labels C$_j$ from the original training data. Validation tuples are selected from the general data set at random. They are independent of the training tuples and were not used to construct the classifier. Then the accuracy of the classifier on the validation dataset is evaluated by the percentage of test set tuples that are correctly classified by the classifier. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known or can be used to simulate data tuples.

B.) Bayes' Theorem

Bayes' theorem (also known as Bayes' rule) is a means for calculating the conditional probability of a random event given some additional information. In this section, it is described how Bayes theorem could be used for classification of random events (i.e., tuple X) of a given database. In this example, X is a tuple of the given database that is identified by measurements

of its attributes. For such case, we assume there exists a hypothesis indicating that the data tuple X belongs to a specified class C. Then, our goal is to determine P(C|X), or the probability that tuple X belongs to class C, given our knowledge and evidence on the attribute description of X. P(C), P(X|C), and P(X) are called the prior probability and are estimated from the a set of statistical evidence. The posterior probability, P(C|X) is calculated using the following equation based on Bayes' theorem:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

As an example, having applied Bayes theorem to issue of correctly diagnosing prostate cancer, if we select men at random from a male population and remove their prostates for definitive diagnosis of cancer we would find the prior probability of prostate cancer, P(*has cancer*), in the given male population. If we want to know the probability that any randomly selected man has cancer without removing his prostate, we would want to know all relevant information such as age and Prostate Specific Antigen (PSA) level. We generate a dataset drawn from a random sample of men that includes data on the patients' age and PSA level. A given patient with the age of 70 years old and a PSA level of 4.0 ng/mL is represented by tuple **X**. Suppose a physician hypothesizes that her patient has prostate cancer. Then the posterior probability or P(*(has cancer)*|**X**), indicates the probability that this random patient will be diagnosed with a prostate tumor given that we know the patient's age and PSA level. In this case we can use Bayes theorem to find P(*(has cancer)*|**X**) when we are given P(*has cancer*), and P(X|*(has cancer)*) or the probability that a prior patient is 70 years old and has a PSA level of 4.0 ng/mL, given that we know the patient will be diagnose with a prostate tumor.

C.) Pattern Learning: Bayes Classifiers, Bayesian Network Classifiers

A Bayes classifier applies Bayes theorem to the classification problem.
Returning to our prostate cancer example, given a random set of variables,
such as sex, income, car model, and age, we seek a model that can predict
the probability of cancer in a randomly selected individual. More generally, the
Bayes classifier defines a model $\mathbf{M}$ to a random set of training tuples D, where
$D = \{(x^{(1)}, c^{(1)}),..., \{(x^{(N)}, c^{(N)})\}$ of size N from a joint probability distribution $p(\mathbf{x},$
$c)$. The Bayesian approach can be described as a non-deterministic
polynomial-time hard optimization problem under a binary loss function that
minimizes conditional risk based on prior probability (Bielza, 2014). The
optimal model is one that seeks to derive the most probable posteriori class
for any given example $\mathbf{x^r} = (x_{r1}, \ldots, x_{rn})$ drawn from the same data source as
the training tuples, or

$$\arg\max_{c} p(c|\boldsymbol{x_r}) = \arg\max_{c} p(\boldsymbol{x_r}, c)$$

In contrast to the heuristic Bayes classifiers, constraint based Bayesian
network classifiers approximate $p(\boldsymbol{x_r}, c)$ according to a Bayesian network. A
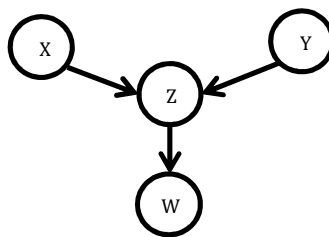
Figure 3.1. Example DAG

Bayesian network is a graphical depiction of a data called a directed acyclic
graph (DAG) where the nodes of the DAG are the c classes of any given
variable $X_1,..., X_n$, and the edges (or vertices) define the (in)dependence
relationships between those variables. In Figure 3.1, the nodes X, Y, Z, W

correspond to the class C within a given data set. The arrows that define the *asymmetrical* relationships between the nodes X, Y, Z and W describe the probabilistic conditional dependences between the nodes. In other words, the prior probability of a given parental node such as X or Y can predict probability that any given tuple belongs to a particular class of child nodes such as Z and W. Once a Bayesian network (i.e., DAG) is constructed, it provides a logical and interpretable framework for learning the probability of any given set of data. The fact that each variable is conditionally independent of its non-descendants in the DAG, given its parents, allows the network to provide a complete representation of the existing joint probability distribution with the following equation:

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i | Parents(Y_i))$$

where $P(x_1, : : : , x_n)$ is the probability of a particular combination of values of X, and the values for $P(x_i | Parents(Y_i))$ correspond to the probability for Y_i based on a set of training data.

There are multiple methods for Bayesian networks to approximate joint conditional probability distributions or $p(x_r, c)$. Naïve Bayes modeling takes the simplifying assumptions that some class C is the parent of all predictor
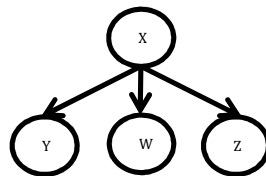
Figure 3.2. Naïve Bayes

variables and there are no independence relationships between the predictor variables. In Figure 3.2 we see that X is the causal parent to Y, W and Z but it is assumed that there is no relationship between Y, W, and Z and it is

assumed that x has no parents. Although naïve Bayes requires very strong assumptions, it has proven effective in several machine-learning tasks (Garcia-Laencina, 2013). However, in many models the simplifications prove too great to effectively model the data. A more complex variant of naïve Bayes is an 'unrestricted' Bayes pattern seen in Figure 3.3. This graphical
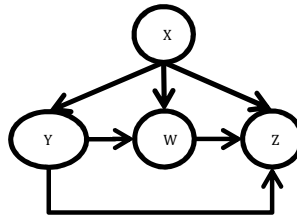
Figure 3.3. Unrestricted DAG

approximation increases the number of conditional relationships between the nodes and thereby dramatically increases the number of states within the Bayes net. Models constructed from unrestricted patterns face the problem of overfitting and subsequent poor parameter learning.

A more sophisticated approach allows that X could have some parent 'a' and there are potential relationships between the variables Y, W and Z. In order to find the causal parents of X we must search the Markov blanket of X. Once we find DAG describing the conditional dependencies between classes within the Markov blanket, we assume this blanket represents the probabilistic dependences existing within a set of data. That is, the behavior of the true DAG equates to the DAG discovered within the Markov blanket. This assumption is known as the faithfulness assumption (Pearl, 1988; Sprites 2000). The problem of classification thus becomes one of relationship discovery within a projection of some tuple X onto its respective Markov blanket.

A key concept in the graphical representation of conditional independence is the concept of *d-separation* (Figure 3.4). This concept first described by (Pearl, 1988) provides the justification for simplified conditional independence relationships within a given Markov blanket. For example, we say that X and Y
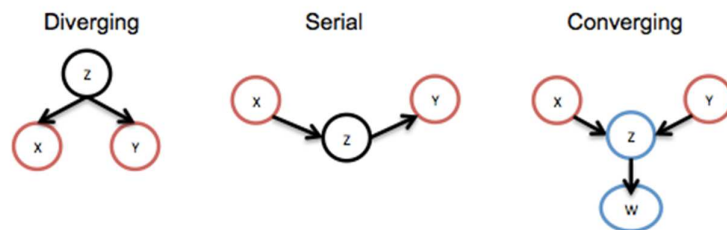
Figure 3.4. D-Separation.

are *d-separated* if on any directed path between X and Y there is some variable Z that such that Z is known and is either a diverging parents connection or is known and is in serial connection between X and Y. X and Y are also d-separated if there is some common unknown child that X and Y converge on, and Z has an unknown descendent. Full descriptions can be found in (Pearl, 2000; Pearl, 2009), while truncated but general description is explained in (Dawid, 2010).

There are multiple methods for discovering the relationships within the Markov blanket and subsequently Bayesian network (i.e., DAG) discovery. Many of these published methods of DAG structure learning have been executed on real world datasets (Kalisch, 2010). Below we give a brief introduction into several different methods for conducting DAG structure learning as well as the assumptions, strengths and weakness. DAG structure learning procedures are usually highly variable, i.e., the learnt graph tends to change drastically with even small perturbation of the data. We then elaborate our method which accommodates the instability of Bayesian search algorithms through a

combination of ensemble learning techniques each proven to individually increase the truthfulness of Bayesian search output.

There are three branches to which we classify DAG learning methods employed on real world data sets, score-based, constraint-based and hybrid methods. Score based methods have essentially two parts, a scoring and an aggregating method that seeks to optimize the decided upon scoring technique. There are several ways of scoring such as negative log-likelihood score, Akaike information criterion (AIC) score, Bayesian information criterion (BIC) score, Bayesian Gaussian equivalent (BGe) score (Geiger, 1994). At each step the search method determines if including, orienting or deleting an edge between two notes will increase or decrease the score and optimizes for the lowest score. Because the potential relationships within a highly dimensional data wrapper are prohibitively large, greedy search algorithms, particularly hill climbing algorithms are employed. Greedy Hill climbing algorithms optimize local relationships in a forward step adding edges until a maximum score is achieved and then a backward step deleting edges until the scoring criteria can no longer be improved (Chickering, 2002).

D.) Markov Blanket Based Pattern Discovery

Constraint based DAG structure learning views a DAG as the result of a set of conditional independence tests applied across the Markov blanket of nodes. As such in contrast to scoring algorithms that approach classifier search as a simplification of the general classifier optimization problem, constraint based DAG searches approach classifier search as a feature selection problem. To select a feature, the data is tested against multiple different hypotheses. There are many published Constraint based algorithms with variety of

different assumptions and potential applications. Several of the mostly widely used algorithms are variants of the PC algorithm (e.g. Conservative PC, JCPC, PCD etc.), named after its creators Peter Sprites and Clark Glymour. Additional algorithms that all reside in this branch include Increment Association Markov Blanket (IAMB) and two IAMB variants (Margaritis and Thrun, 2000). There is also a large family of constraint-based methods designed to deal with latent variables (i.e., variables not found within the dataset). These methods, FCI, FCI variants, IDA and several others methods all output Partial Ancestral Graphs (PAGs) that can determine if measure variables could be the result of some unmeasured (latent) variable.

Lastly, a third type of popular Bayesian search algorithms can be considered combinations of the score and constraint based DAG searches, also called hybrid methods. In hybrid methods, conditional independence tests are used to determine edges, but each local test between nodes are used to inform the proceeding tests. Primarily, the knowledge from each search is used to impose restrictions on the search space via a scoring system. Hybrid methods include Max-Min Hill Climbing (MMHC) (Tsamardinos, 2006) and L1MB (Schmidt, 2007).

Constraint and Hybrid search algorithms can commonly be broken down into two phases, the search phase and the orientation phase. During the search phase, the algorithm asks a conditional independence oracle to perform a routine statistical test, usually either χ^2 contingency test or G-test to determine if a pair Markov equivalent nodes are independent or not. Within the pattern discovery catalogue of algorithms, there are additional statistical methods for hypothesis testing that may be better suited for a dataset. Just as with any χ^2

or G-hypothesis test, the user has the ability to set an α-value or significance value, above which the null hypothesis is rejected. These conditional independence tests, and therefore the Constraint based and hybrid searches demands that the data be either entirely continuous or entirely discrete.

The PC family of algorithms performs a backward stepped algorithm that begins with a maximally connected Markov blanket. Relationships between the classes within the Markov blanket are then tested. If the statistical test rejects the null hypothesis the edge between two nodes in the Markov blanket is maintained. If the test determines the null hypothesis true, the edge is deleted stepwise until a graph depicting all true relationships. The conditional independence oracle describes a strategy for passing the tuples through the series of conditional independence tests that are performed across the Markov blanket.

If two Markov nodes reject a null hypothesis via the conditional independence test, the algorithm outputs an undirected edge between two variables. Once all necessary independence tests have been applied across the Markov blanket, the algorithm enters the orientation stage. Here the algorithms determine directionality of the edges via application of the *d-separation* principle across triplets of tuples as well as the orientation rules described by Meek (1995). Orientation of the arrows is particularly susceptible to small perturbations of the data and can result in a partial failure of the algorithm. Partial failure would include multiple undirected edges and edges that are bi-directional. When search results demonstrate such a partial failure, a Pattern Graph can include several symmetrical relationships. Since DAG represents a collection of asymmetrical relationships, a single search result can suggest

multiple potential DAG's.

E.) Parameter Learning: Bayesian Belief Network

A Bayesian belief network (G, Θ) is composed of two elements. 'G', a directed

acyclic graph (DAG) and 'Θ' a subsequent collection of conditional probability

tables (CPT) defined by the Bayes net parameters draw from the DAG. Figure

(a)

CPT P(LC | FH, S)	FH, S	FH, ~S	~FH, S	~FH, ~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

(b)
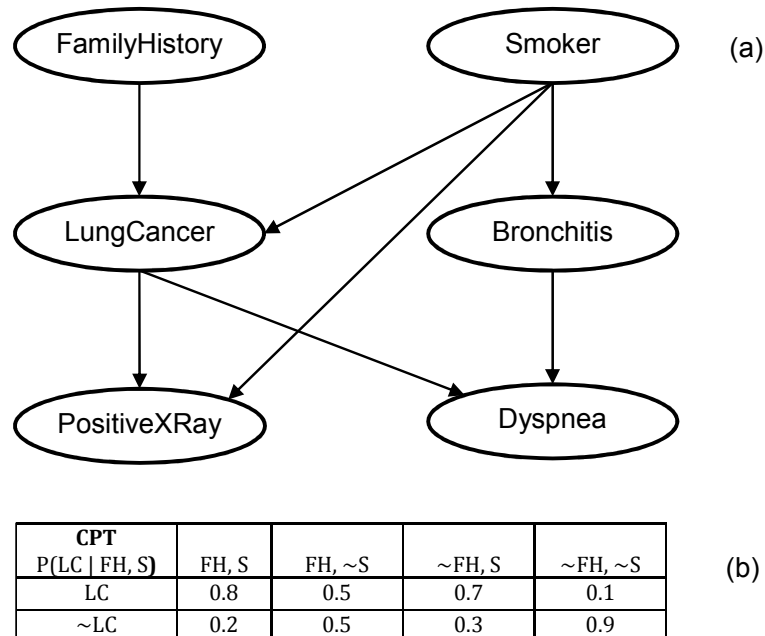
Figure 3.5. A simple Bayesian belief network: (a) A proposed causal model, represented by a DAG. (b) The conditional probability table (CPT) for the values of the variable LungCancer (LC) showing each possible combination of the values of its parent nodes, FamilyHistory (FH) and Smoker (S). Figure is adapted and modified from (Jiawei Han, 2006).

3.5 depicts a modified Bayesian belief network example from (Jiawei Han et al

2006). The DAG defines certain states within a bayes net, the parameters of

which can then be learned. The DAG and subsequent parameters may

correspond to actual attributes given in the data or to "hidden variables"

believed to form a relationship (e.g., in the case of medical data, a hidden

variable may indicate a syndrome, representing a number of symptoms that,

together, characterize a specific disease).

The belief network in Figure 3.5 includes six binary variables with the ability to infer the probability of either a '*PositiveXRay'* or '*Dyspnea'*. For example, having lung cancer is influenced by a person's family history of lung cancer, as well as whether or not the person is a smoker, but is d-separated from bronchitis. Note that the variable '*PositiveXRay'* is independent of whether the patient has a family history of lung cancer or is a smoker, given that we know the patient has lung cancer. In other words, once we know the outcome of the variable '*LungCancer'*, then the variables *'FamilyHistory'* and '*Bronchitis* do not provide any additional information regarding '*PositiveXRay'*. That is, given the rules of d-*separation*, p(*PositiveXRay | Smoker*, *Lung Cancer)* and p(*Dyspnea | Lung Cancer, Bronchitis).* The edges also show that the variable '*LungCancer'* is conditionally independent of '*Bronchitis'*, given its parents, '*FamilyHistory'* and Smoker.

Thus, a DAG provides a probabilistic Bayes net that approximate the optimal inferred probability. A belief network includes one conditional probability table (CPT) for each variable. Any node within the network can be selected as an
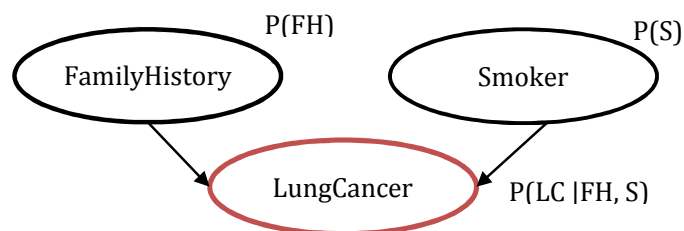
Figure 3.6. An Example of CPT for the variable "*LungCancer*".

"output" node, representing a class label attribute. There may be more than one output node. Figure 3.6 includes an example CPT for the variable '*LungCancer'*.

The CPT, labeled (b) within the Figure 3.5, provides the conditional probability for each known value of '*LungCancer*' is given for each possible combination of values of its parents.

P(*LungCance*r = yes | *FamilyHistory* = yes, *Smoker* = yes) = 0.8

P(*LungCance*r = no | *FamilyHistory* = no, *Smoker* = no) = 0.9

The probabilities that satisfy the specified states within the Bayes net are called 'parameter learning'. There are several algorithms for learning, or estimating, the parameters, Θ, of a network. When the data is complete (e.g., without missing values), the Maximum Likelihood Estimator, MLE can be used to calculate the conditional probability of a given parameter, where D is any given node within the network

$$\max_{\Theta} \log P(D|\Theta)$$

This limit of which is calculated via the following, where N_x is the number of counts within the training data set as defined by **pa**$_x$ or the parents of the desired node**.**

$$\mathrm{ML}(\theta_{x,pa_X}) = \frac{N_{x,pa_X}}{\sum_x N_{x,pa_X}}$$
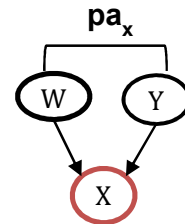
A similar version of the MLE is called the MAP-estimator that relies on a dirchlet distribution of variables. This allows the user to specify with prior evidence or knowledge the specific distributions or priors of the model.

$$\max_{\Theta} \log P(D|\Theta)P(\Theta)$$

$$\text{MAP}(\theta_{x,pa_X}) = \frac{N_{x,pa_X} + \alpha_{x,pa_X}}{\sum_x N_{x,pa_X} + \sum_x \alpha_{x,pa_X}}$$

Both the MLE and the MAP-estimator demonstrate similar behavior; they both are asymptotically equivalent and consistent as the counts are varied within the training model. More importantly both algorithms rely on *sufficient statistics*. The concept of sufficient statistics is two fold, both of which are critical for properly training the parameters of a model. First, if there is no training data provided to estimate the likelihood of a given parameter the algorithm has no sensible way for calculating the conditional probability. These cases are most common in data with significant outliers or subclasses within the data.

Second, sufficient statistics also implies a subtle rule that assumes there is no bias within the original training data set. For instance, if we return to our example Bayesian belief network at the beginning of this section we can imagine that if we are to sample only individuals that were already in the hospital instead of the general population, our maximum likelihood estimation would primarily consist of individuals that did not represent the general representation of the belief network. Unintended bias is a significant contributor to weak and ineffective Bayesian modeling.

If the data has missing values the ML and MAP algorithms are unable to produce estimates. In these cases, parameter learning must be accomplished via the Estimator Maximizer (EM)-algorithm or some similar variant that has a sensible means of dealing with the missing data. The EM-algorithm is a two-step algorithm. In the first part, it computes an expected count missing value based on inference from the Bayesian network.

$$E_{P(X)}[N_{x,\boldsymbol{pa_X}}] = \sum_{k=1}^{N} p(x, pa_X \mid y^k, \Theta, G)$$

Once it computes the value it estimates the probability for the given parameter using the ML or MAP algorithms. This continues iteratively until the probabilities converge within some predetermined threshold, usually 0.0001. When large amounts of data are missing or when multiple latent variables exist, learning parameters becomes increasingly variable. Additionally, it has been repeatedly demonstrated that since the EM and similar algorithms have weaker guarantees than the ML or MAP algorithms, they become trapped in local maxima (Liao, 2009). Although there are several modified versions of the EM algorithm that have improved on the original presentation, as of yet, none provide robust results when missing values for any given parameter rise above 30% (Kohavi, 1999). Therefore, domain knowledge, and data preprocessing retain critical importance when constructing the training data to provide the model.

Method

A.) Bayesian Network Modeling for Clinical Avatars

Stoll and Schubert (Keeler, 2006) suggest a four step semantic chain from raw data to wisdom. Similarly, our method follows a four-step logic chain to progress from a set of data to clinical avatars (Figure 3.7). The method can be
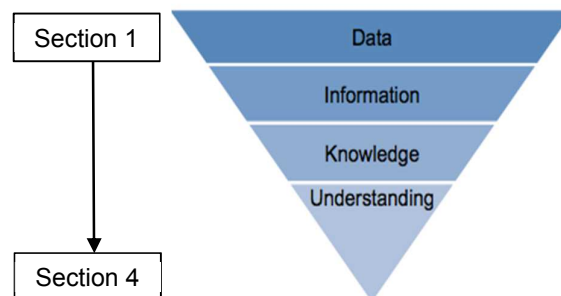
Figure 3.7. Semantic chain from raw data to understanding.

partitioned into four broad sections: (1) Data and knowledge aggregation and preprocessing. During this step, the data must be acquired from a source and characterized alongside any additional information regarding the nature of the collection method and data dictionary that could unintentionally bias the resultant clinical avatars. Additionally, and expert knowledge from published or unpublished sources should be developed and employed to enrich the patient data. (2) In the second section, an ensemble of Bayes search algorithms are employed along with the domain knowledge acquired in section 1 to discover any significant relationships between variables present in the patient data. The relationships are then mapped graphically in one or several Directed Acyclic Graphs (DAGs). (3) In section three the DAGs are used to construct an ensemble of parametric models that allows the estimation of joint conditional probabilities and subsequent deriving an instantiated model. Once the ensemble DAGs and conditional probabilities have been aggregated into data generating models we consider it a Bayesian network model (BNM). (4) In section four, the BNM is cross validated against a subset of the original data held out from sections 2 and 3. Each section therein consists of several steps summarized in Appendix Figure 3.21.

Section 1: Data Preprocessing and Gathering of Domain Knowledge

Before modeling can begin, the appropriate data must be aggregated into a data wrapper. The method described here is flexible. Almost any type of healthcare data can be used to train the BNM. However, in accordance with the "No-Free-Lunch theorem", (Wolpert, 2008) minor modifications in the pipeline to accommodate specific data sets will invariably benefit the model. The quality of the data used to train and validate the BNM will reflect almost

identically the quality of the BNM produced. It is quite possible to produce a perfect model from imperfect data, a model with little relevance. Therefore, despite the empirical principles this method employs, domain and expert knowledge remain critical. Further, it is paramount that care and thought be applied during the querying and accumulation of the data before embarking on the modeling process. Once this is accomplished, the data can be brought into the pipeline for generating clinical avatars.

This section has two branches performed in parallel. In Part I we accumulated domain knowledge (i.e., expert knowledge or literature-defined knowledge) to better understand both the specifics of our data such as the way the data was gathered, the semantics of the data dictionary and any measurement error as well as the general relationship between the variables as found in literature review. In Part II the patient data is characterized and prepared (i.e., data preprocessing). Data preparation is a multistep process as described in Appendix Figure 3.21. Once the data has gone through the data preprocessing procedure it is aggregated into a data wrapper.

Part I:

It has been said, "there can be nothing fully automatic about causal discovery" (Dawid, 2009). Domain Knowledge, both regarding the nature of the sample, the techniques involved in imputing or deriving the data and the general relationships between the variables is all critical for the development of clinical avatars. Domain knowledge is critical in accurately preparing the data for modeling. Deciding how to aggregate the data, how to accommodate outliers and appropriate handle missing values are all depending on the researcher

making explicit choices about the nature of the data and the intent of the final model.

Beyond the data preparation, domain knowledge performs three definite functions within the clinical avatar pipeline. First, domain knowledge can constrain the search space of the Bayesian search algorithms by imposing known relationships across the data. The search space can be constrained via, required causal relationships and forbidden relationships. Required relationships are defined by a corollary link between variables previously established in peer-reviewed literature. Forbidden causal relationships restrict the search from finding false positive causal relationships in the data. The variables can also have a series of relationships defined by hierarchically categorizing the variables according to the principles of causation. By classifying the variables in this manner, the user forbids directed relationships from higher to lower tiers. Constraints typically make the structure learning more efficient and can improve the validity of the resultant DAG.

The second function of domain knowledge is to provide directionality to the arrows across the Bayesian search results. Arrow directionality within a DAG is a subtle and nuanced subject, the significance of which is discussed in an upcoming section. Often times Bayesian classifier algorithms produce ambiguous directional results. There are three sources for this ambiguity because of the weak guarantees within the proofs necessary for theorems to be true, uncertainty within the training data and the inherent weak philosophical underpinnings of causality in the first place. In fact, some statisticians view causality as nothing more than a convenient concept. As an additional confounder, there are certain times when the directional causal

relationship may change over time. To give a short example borrowed from Sprites (2000), if the variable under consideration is the rotation of a car tire, I could determine that the function of the car engine is causal parent of the tire rotation. However, if I were to try and kick-start the engine via pushing the car down a hill and use the motion of the car to start the engine, then the direction of the rotation of the tire would be the causal parent to the engine starting. The result is often a direct *cyclic* graph and the directionality of certain causal relationships must be determined via domain knowledge. It is important to note these ambiguous or time dependent relationships as best as possible within a given dataset.

Domain knowledge also comes into play while aggregating the Bayes pattern learning results. Once the results are aggregated, we use domain knowledge in parallel with empirical methods to determine the edges that should most likely be pruned. In general, domain knowledge can help support the simplifying assumption of developing DAG while also preparing the data to reduce the violations of those assumptions. In all three points of applying domain knowledge ad hoc to the modeling method there is a varying amount of uncertainty. However, to account for such inherent uncertainty in the model we combine several plausible causal structures into the final pipeline.

Part II:

Data Reduction and Characterization:

Here the variables that are desired to be included in the model are segregated from the dataset. This includes the establishment of a data dictionary to document the significance of each variable. The goal here is to statistically describe the nature of each variable following data parsing. It is important to

pay attention to the outliers and any missing values within the data.

Data Cleaning:

Of particular important in this process is the role of missing values. If there are no missing values, then we proceed to "Data Discretization and Aggregation". Missing data can compromise the robustness of any statistical model and additional steps must be used to handle missing values. First the nature of the missing values must be determined. There are three forms of missing values, 'missing at random', 'missing completely at random' and 'not missing at random'. Missing at random denotes cases where missing values are scattered around the data at random and there are no hidden variables that contain missing values. 'Missing complete at random' denotes datasets where there is a hidden variables that is not represented in the training data and not missing at random where all variables are present, but missing values are concentrated in local classes.

1.) If the missing values are randomly distributed and deleting all patient cases with missing values does not significantly alter the representativeness of the data, then delete the all rows of missing data and proceed to "Data Discretization and Aggregation". We define significance as a change in proportion of 30% or greater in any single variable (Friedman, 97; Ramoni, 2000).

2.) If the missing values are randomly distributed, but deleting the patient cases significantly alters one or more variables, then proceed to the following step.

3.) If the missing data is non-randomly distributed or if the missing values are randomly distributed, but deleting the patient cases significantly

alters one or more variables, then imputation should be considered. The specifics of imputation are outside the scope of this document. If an evidence based method exists for imputing missing values that the research is confident in, the researcher should apply those methods wherever possible.

4.) If the imputation method was capable of increasing data coverage to >70% in any given variable then proceed to "Data Discretization and Aggregation".

5.) If following execution of the imputation method, if any variable has more than 30% of values missing, then randomly assign values until data coverage increases to >70%.

Every effort should be made to ensure that missing values are treated in the appropriate method. Additionally, not all missing values are of equal importance within the DAG it will train. Variables that are discovered to be the parent nodes to other child nodes are more sensitive to missing values than variables in the child nodes. Deleting data with missing values can both introduce bias into the results and prevent accurate estimation of conditional probabilities with the training the data. Each data set must be prepared according to the specific characteristics of the data. Preparation of the data should be approached as part of the experimental process in developing clinical avatars.

Data Discretization and Aggregation:

Often healthcare data includes a mix of categorical and discrete data, such as race and gender, and continuous data such as age and height. The Bayesian search algorithms we employ require training data that is either entirely

continuous and normally distributed or entirely discrete. The Aurora Health Care dataset included a combination of both continuous and discrete variables, therefore the continuous variables were discretized. The process of discretizing continuous data is a ubiquitous data preprocessing technique that must balance information loss inherent in the process with the benefits of greater processing efficiency. There are numerous discretization methods and the choice can impact both the posterior probability estimation as well as the discovery of inherent causal structure in the underlying graph. We employed a common unsupervised method, EqualWidth that has demonstrated its ability to produce accurate data mining results for Bayesian search algorithms when compared to other techniques (García-Laencina, 2013).

Generally, there is a balance between information loss inherent in the discretization process and the need for sufficient population sizes to estimate conditional probabilities. For example when EqualFrequency is applied to normally distributed continuous variables, datums that approach the minimum and maximum are grouped into bins with more frequent ages. This has the advantage of providing additional data to determine the conditional probability of any particular state. In contrast, Equalwidth provides age bins that are consistent in size but have some bins with low frequency. In regard to learning the network structure, Sprits noted that a distinct challenge in discretization of continuous features is that conditionally independent continuous variables may be transposed into non-conditionally independent discretized foils. Therefore it is important to evaluate discretization methods that maintain the underlying causal relationships (Sprites, 2000). In practical application with the Aurora data, we found consistency and improved performance in

Bayesian classifier search results using EqualWidth when compared to other methods.

Section 2: Directed Acyclic Graphs and Ensemble Learning

Weakness within Bayesian Classifier searches:

As outlined above, there are many Bayesian techniques for learning the causal structure of the variables. However, each of these methods is built upon several assumptions. For example, the PC algorithm and associated PC variants (e.g. CPC and JCPC) have the following assumptions: No hidden or selection variables that would suggest the number of variables grows with the sample size, if the underlying DAG is sparse, the data is multivariate normal and satisfies some regularity conditions on the partial correlations (Kalish, 2007) and (Kalich, 2014). In contrast the FCI family of algorithms presumes that there may be hidden and selection variables; consistent in high-dimensional settings if the so-called Possible-D-SEP are sparse (Sprites, 2000), the data is multivariate normal and satisfies some regularity conditions on the partial correlations. These assumptions are necessary for the truthfulness of these search algorithms, yet they can often be violated and produce accurate results (Sprits; 2000; Domingos, 2012). How or to what degree the search assumptions may be violated and remain truthful is not consistent between data sets. Additionally, there is no known algorithm for determining the bounds that any given algorithm can be violated.

Sprites (2000) does however offer a general guide of nine factors that determine the precision and accuracy of a DAG: 1. The correctness of the background knowledge, 2. How closely the Causal Markov Condition holds (e.g., no inter-unit causation, no mixtures of subpopulations in which causal

connections are in opposite directions), 3. How closely the Faithfulness Condition holds (e.g., no deterministic relations, no attempt to detect very small causal effects), 4. Whether the distributional assumptions made by the statistical tests hold (e.g., joint normality.), 5. The power of the statistical tests against alternatives, 6. The significance level used in the statistical tests, 7. The sample size, 8. The sampling method, 9. The sparseness of the true graphical model. Our goal then is creating a method that can be applied to any healthcare data set is to produce an evidence based method that can address the weaknesses in each Bayesian learning algorithm while simultaneously optimize the nine above conditions to determine the precision and accuracy of the DAG. In the following section, we describe several ensemble-learning techniques used in concert for increasing the robustness of DAG learning when employed on real world data sets.

Bayesian Ensemble Learning

Although there may be one particular algorithm that performs better for one data set over another, instead of experimenting to find one particular superior variation, the researcher can include many different algorithms at once with little additional effort. In many branches of applied machine learning ensembles have become the standard. As computational power increases alongside new combinations statistical techniques already developed (as well as those underdevelopment) the trend is towards ever-larger ensemble techniques. There are many techniques that have demonstrated improved results by linearly combining several Bayesian classifiers into an ensemble technique. The question then becomes what combination of ensemble techniques can best be applied to develop clinical avatars.

There are three widely tested and proven techniques within Bayesian ensemble learning, bootstrap aggregation (Bagging), Bayesian boosting and stacking. Bagging is often considered the simplest technique. From one set of training data, the data is resampled at random and from each resample passes through a Bayesian search algorithm. Bagging has been shown to limit increase bias while dramatically increasing variance. There are both multiple methods for generating the random bootstraps and for combining results of the Bayesian searches. Boosting builds upon a similar principle as bagging, however, the results of each classifier result has a weight. The weights for each training set are varied so the proceeding classifier, training examples have weights, and these are varied so that each new classifier focuses on the previously weak results. In essence, boosting using several weak Bayesian classifiers and combines them into a single strong Bayesian classifier. In stacking, the outputs of individual classifiers are feed as input into a second Bayesian classifier. The second Bayesian algorithm then decides the best way to aggregate the results (Domingos, 2012). Our method includes the following ensemble techniques, bagging, two distinct methods of constraining both score and constraint based search algorithms with of domain knowledge, and lastly Bayesian Model Combination.

The use of multiple bootstraps has been shown to address bias and variance reduction within the data and therefore reduce the number of false positives and false negatives (Friedman et al 1999). The demonstration of why bagging works and its implications has been discussed at length in previous publications (Domingos, 1997). As noted above, bagging generates an ensemble of DAGs from the bootstrap resamples of the training data and then

minimizing the overall distance to the entire ensemble derives an aggregated DAG. The size, the number of variables and character of those variables has an important role to play in the method used to bootstrap and segment the data into validation and training subset. There are several methods for bootstrapping that have demonstrated optimal results for particular sets of data. Four of the most common bootstrap methods were compared in a review by Broom (2012); the classic with replacement bootstrap, the Bayesian (or parametric) bootstrap, the bias corrected bootstrap and the double bootstrap aggregation (i.e., bootstrapping a bootstrap). Generally speaking the Bayesian bootstrap and the classic with replacement bootstrap perform comparably. There are several conditions that inform the researcher as to which technique should be applied to which data set (Broom, 2012). Bagging averages out General Noise features that cannot be cleaned from the data over progressively larger ensembles. Because we employ a pruning technique that removes edges with less that 50% commonality, this bagging also reduces false negatives.

In experimentation with EMR data we employ a "repeated leave one out bootstrap aggregation" method (Clyde, 2004; Jiang 2007). Holding out data for cross validation is considered a data mining 'best practice' (Belazzi, 2004). However, when the sample size is small and outliers are important considerations in the data, holding out any amount of data reduces variance in the model. Therefore, it is recommended to perform the bootstrap sampling of the original data prior to the dividing the data between training and validation subsets. The choice of bootstrap sampling technique can ultimately impact the quality of the model.

The bootstraps should be drawn randomly with replacement to subsample of equal size to the original data. Regardless of what technique is used for bootstrapping, once the data is bootstrapped, then each bootstrap should be divided into training and 20% validation subset. The validation data should be pooled for validation of the BNM, while the training subsets should remain distinct for DAG training.

The question of how many bootstrap resamples are needed is not easily answered because it depends on the specific sample size, dimensionality and characteristics of the data. Previous studies (Friedman, 1999) have used anywhere from one to two hundred bootstrap resamples to 2500 bootstrap resamples (Broom, 2012). No experiments to date have determined how many bootstraps are necessary when using conditional independence Bayesian search algorithms (such as the PC algorithm) with multidimensional data. However, our results demonstrate some consistency with the asymptotic improvement of search results with minimal increase in bias. Generally, the number of bootstraps should increase as the sample size gets smaller and the dimensionality gets larger. Across several types of Bayesian search classifiers, the ability for bagging of any variety to correctly infer causality breaks down between n=125 and n=250. We found that in healthcare data with 20 variables can be resampled with as few as 5 bootstraps and significantly increase variance and improve search results.

As an additional method for increasing variance and correcting for bias in the generation of clinical avatars we recommend performing an ensemble of Bayesian classifier search algorithms on each bootstrap of data. Although it is likely that one search algorithm will produce the most accurate DAG, by

performing multiple types of Bayesian classifier searches on each bootstrap we found iterative convergence on particular edges. There are multiple ways to increases variance and adjust for bias in the selection of Bayes search algorithms. Since each particular Bayes search algorithm has specific assumptions (e.g., PC algorithm vs. IAMB) and it may be unknown as to which assumptions best suit the data, we suggest performing several types of algorithms on the same dataset. For example, given a set of training data the CFCI may produce superior results to the CPC algorithm if the data contains latent variables that affect the conditional probabilistic dependences since the CFCI does not make the same assumptions of no latent or hidden variables perturbing the data. If however there are known latent variables but it unknown definitive whether they affect the probabilistic dependences in the data, than the CPC algorithm may produce superior results.

If the data has significant non-random missing values, or other forms of sampling bias, we recommend a second method for broadening the ensemble of Bayes searches. This method involves constraining some search algorithms with domain knowledge and allowing other the search algorithms to perform the search unconstrained the adjusting the orientation of the edges according to some predetermined tiers (see Section 1 Part II). The PC algorithm as well as variants of the PC algorithm (e.g., CPC, JCPC) is receptive to constraining the search space via domain knowledge. If you impose particular forbidden or required edges on the search space, the conditional independence oracle will adjust the series of independence tests it performs across the data. In highly dimensional data sets with limited sample size, this can increase the variability between search results. An additional

method involving ambiguity or time sensitive uncertainty within the domain knowledge is to modify the domain knowledge between certain bootstraps or groups of bootstrapped data sets. This enriches the hypothesis space that allows for certain DAGs to be true at certain moments in time and allowing others to be right at other times. As similarly recommended by Spirites (2000), we do not recommend the Bonferonni adjustment for multiple hypothesis tests. However, if perform a large number of searches on the same data vector, it is possible to increase α to a more stringent parameter such as α = 0.01 or 0.001. In respect to the PC family of algorithms, the graphical output because increasingly sparse as α value increases.

Once all the necessary searches have been performed on the data, the resultant DAGs must be aggregated. There are several methods described for combining Bayesian search results. In accordance with other applications of Bayesian ensemble learning, we recommend combining all edges for each bootstrap and then pruning the edges according to some cut off. Depending on the dataset, some have pruned edges that were returned in greater than 20% of search results, while others have selected edges that received 50% or more commonality. As of yet, there is not demonstrated empirical technique to determine how to aggregate edges within an ensemble of Bayesian search results. Therefore, we recommend applying both the 50% common edges and/or those edges that are supported by the domain knowledge aggregated prior to performing the search.

With any data or domain knowledge that has the potential for significant bias, we recommend performing Bayesian model combination. Bayesian model combination has been demonstrated to improve search results when

compared to and combined with other ensemble learning techniques, (Monteith, 2011). Bayesian model combination is unique from aggregating multiple search results into one DAG because it allows for multiple potentially true DAGs to exist given ambiguous training data and or domain knowledge. Bias can be significant if the original training data contained missing values approaching 30% in any given variable and/or there was imputation performed. The essential element of Bayesian model combination is that instead of converging all Bayesian searches performed in this section into a single DAG, several resultant DAGs are then used in parallel for section three. Each DAG is used to derive a parametric model that is input into conditional probability estimators. Each DAG plus conditional probability will ultimately generate data via Monte Carlo simulation. The data is then aggregated via either a Dirichlet weighted distribution or an unweighted distribution. Lastly, to further demonstrate the robustness of particular classifiers, this aggregation of data can be searched and the results compared to the DAGs used to model the input data.

A strength and a weakness of DAG's are their openness to interpretation. No fewer than half a dozen different interpretations are used in the application of DAG's to various fields and in their theoretical discourse. Standard practice in Epidemiology dictates that the directed arrows within the DAG have direct causal meaning, (e.g., "X → Y" would read as X is the causal parent to Y, or even stronger, X is the cause of Y) (Evens, 2012). There are three explicit assumptions that must be undertaken to assume that DAG's represent a causal map: 1.) There exists some true DAG that is a causal DAG representation of the system being studied, 2.) This causal DAG is identical to

the DAG representative of the probabilistic conditional independence properties of the system, 3) The third is the faithfulness assumption, that the causal DAG is a probabilistically faithful representation of the system. A weaker interpretation has been described as probabilistic conditional independence of two particular variables and the causal independence of those same two variables given a third (Sprites, 2008).

In the development of clinical avatars we apply the most robust and theoretically truthful interpretation of the DAGs. According to a strict theoretical presentation, A DAG mirrors the symmetric conditional independence relationships within the data and imposes asymmetric probabilistic relationships between the variables via geometric interpretation. The directionality of the arrows, in a purely theoretic plane, has been described as "artifact…although it plays an important role in the formal syntax of the model it has no direct counterpart in the world and contributes only indirectly to the semantic interpretation of the model" (Dawid, 2010). Because of the ambiguity inherent in any given DAG, the bootstrap aggregation ("Bagging") procedure provides more robust clinical avatars. By widening the hypothesis space to include a greater collection of possibly true results (e.g., DAGs), we allow for the final clinical avatar population to be maximally consistent with conditionally independent relationships within both the data and the real world representing the data. Any additional reference to specific strong causal relationships found within the DAG search results are strictly outside the scope of developing clinical avatars. We do recognize that this method could be used to guide a more causally complete interpretation of a given set of data. Applying the clinical avatar pipeline would be useful in

inferring causal patterns in a number of fields and datasets. Nevertheless for the development of clinical avatars this strong assumption is not necessary. Rather it is more important to find the significant probabilistic relationships that need to be replicated in order to capture the essence of the data. The resultant DAG from the executed searches is used to derive a parametric model that is input into conditional probability estimators. The DAG plus conditional posterior probabilities will ultimately generate clinical avatars.

Section 3: Estimating Conditional Probabilities via Ensemble learning

Once the DAG is developed in section two there are a variety of computational efficient Bayesian algorithms that estimate the conditional probability of each particular state within the DAG. However, as noted in the introduction section, there are several important assumptions that should be examined before applying these algorithms to the training data. Again these assumptions revolved around statistical sufficiency. We found that bagging in the same manner as completed in section 2 for structure learning, can dramatically improve the resultant clinical avatars. The original training data can be bootstrapped via a parametric bootstrap or a classic bootstrap. However, of critical importance, we recommend again holding out a portion of the data within the bootstrap to be used for validation of the model.

Even with bagging there are challenges of statistical sufficiency. There are times when the available training data is not sufficiently large to calculate the probability of a parameter within the Bayes net. When such an event occurs there are several ways to refine the model to appropriately accommodate the available training data. We suggest that the DAG be simplified via removing edges from the DAG results one at a time in order of confidence until

sufficient cases exist within training data to calculate conditional probabilities. There are several methods to determine robustness of an edge within a DAG. We recommend the following logic flow for pruning edges from the DAG until there are sufficient cases in the training data to calculate the conditional probability for all relevant states from the parent nodes.

1.) If ensemble Bayesian searches were performed on the data, delete the edge that returned the fewest search results

2.) If there were not multiple searches performed, iteratively complete the search at a more stringent α value until an edge is deleted.

3.) If all search results produce the same causal parents to the point of $\alpha = 0.0000001$ then use evidence and or domain knowledge to delete the least robust edge

4.) Lastly, if there is no relevant domain knowledge or unambiguous evidence then use regression to delete the least robust edge.

Once the DAG has been simplified, the conditional probabilities derived from the simplified model were then used as input for the original parametric model.

The second element of statistical sufficiency is the issue of missing values. Although as noted in the background section, MLE and MAP-estimators produce stronger results because they are not subject to becoming trapped in local maxima within high-dimensional datasets, they are unable to accommodate missing values. Therefore if data has non-random missing values, and certain parameters have a high concentration of missing values while others do not have missing values, we recommend that the parameters without missing values be estimated use MLE or MAP-estimators while the

parameters with missing values use the EM or EM-variant algorithm to estimate the parameters. If any sub-parameter within a categorical data set is also subject to a concentration of missing values greater than 30%, we recommend using the Median probability for the variable on as a whole if binary or the average if it is categorical.

Once the parameters of one or more Bayes net are estimated via the training data, the estimations should be aggregated into a single instantiated model. Here the conditional probability estimates from several algorithms can be combined into an ensemble for each bootstrap of data. Each parameter should be estimated using the most robust technique given the limitations of the available data. Once each parameter is associated with a conditional probability, a set of data can be generated using Monte Carlo simulation techniques. Each bootstrap and instantiated model should have an equal amount of data generated. We have called these sets of data "preliminary avatars" since they satisfy the conditions of being simulated patients. But since the model has not yet been validated they cannot yet be considered "functional" clones of the original patient training dataset. Each set of preliminary avatars should be pooled into a signal set of data.

At this point the pooled preliminary avatars may enter final section for validation. If the preliminary avatars lack a variable, either a latent parent or hidden descendent, the pooled preliminary avatars should enter a final series of parameter learning. Here the DAG(s) already developed from section 2 are again put to use to derive the necessary parametric models that establish the Bayes net. The pooled preliminary avatars can then been used to again estimate the conditional probabilities across the Bayes net. At this stage, only

the MLE or MAP-estimators should be used since there are no missing values within the dataset. If however there is a similar problem with insufficient statistics to estimate the conditional probabilities, a similar 4-step strategy should be used to simplify the conditional dependences within the DAG. This single Instantiated model can then be used to generate the avatars for section four.

Section 4: Validation of the BNM

Following the pattern learning as well as the parameter learning, the model is nearly complete. In section four, the process enters the validation stage of the method. There are many potential ways to validate the BNM. We suggest that following data mining best practices, a portion of the data should be held out for validation. As noted in sections 2 and 3, each bootstrap should have a portion held out for validation. Validation can proceed via aggregating the bootstrap portions of the data, or the validation sections can be maintained distinct and to allow for k-fold validation testing. If the validation data is pooled, there are several methods for validating the model against the pooled validation data. Two of the most basic styles of comparisons used to ensure that the simulated population is representative of the original data. The first is univariate distribution, or the comparative frequency of a single attribute between the training data and clinical avatars. The second comparison is the bivariate frequency distribution between the training and clinical avatar data sets. Variables in which the Bayesian algorithm determined a causal connection and are "d-connected" are plotted in frequency histograms. Different statistical tests should be used to determine variance between the

variables. Z-tests can be used for continuous variables and χ^2 contingency test or G-test can be used for discrete variables.

The development of clinical avatars is an iterative process that often requires perfecting. If significant variance is found in either the univariate or bivariate distribution, the process is refined until there is no significant difference. At each section there are several options that may be refined depending on the type, size and quality of data used to train the Bayesian model. Of critical importance is the use time and care spent in section one, preprocessing the data and aggregating domain knowledge. Different styles of data aggregation and reduction as well as imputation can dramatically change the structure and output of a Bayesian Network Model. If any significant variation is found in the first model, the entire structure should be dissolved and the method should be restarted from section one.

If the clinical avatars have demonstrated no significant variance, then at this step, the latent or hidden variables should be added to the final instantiated model in section 3. The variables that are imposed on the Bayesian network model should be associated with some additional evidence that demonstrates a probabilistic conditional dependency to a variable(s) existing within the model. Thus, when the parameters for the additional variable(s) are imposed on the model the estimated conditional probabilities can be imposed on the Instantiated model. Once all the desired variables exist within the DAG, and the relationships within the Bayes net are associated with conditional probabilities, the BNM is complete and ready to produce an unlimited number of clinical avatars via Monte Carlo methods.

3.2.) Clinical Pharmacology

Anticoagulation agents do not directly affect established thrombi. In the case of occurrence of a thrombus, anticoagulants are usually administered to (a) prevent the growth of the existing clots and (b) prevent the movement of the clots which might result in serious and possibly fatal thromboembolic complications.

Warfarin or coumadin is an anticoagulation agent which inhibits the synthesis of vitamin K-dependent coagulation factors (i.e., Factors II, VII, IX and X).
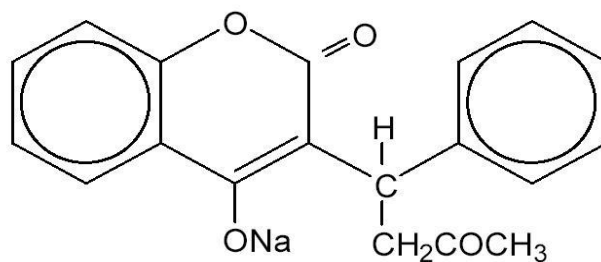
Figure 3.8. The structural formula of crystalline warfarin sodium.

Crystalline warfarin sodium (3-(a-acetonylbenzyl)-4-hydroxycoumarin) is a racemic mixture of the *R*- and *S*-enantiomers. Its empirical formula is $C_{19}H_{15}NaO_4$ (Figure 3.8).

Warfarin is the most commonly prescribed anticoagulant in the world (Pirmohamed, 2006). The number of dispensed warfarin prescription has reached to 30 millions and over 2 million patients are on warfarin in the United States to prevent and for the complications that may occur from thromboembolism (e.g., stroke, heart attack) (Wysowski, 2007; Guyatt, 2012). Warfarin is an effective medication but also has some clinical shortcomings. For instance, a large number of medications and foods interact with it. Many

commonly used medications interact with warfarin. Consequently, it is monitored by international normalized ratio (INR) blood tests to ensure safe adequate doses are taken (Ansell, 2008). An INR beyond the targeted range predisposes to a high risk of bleeding, while an INR below the therapeutic target range indicates that the dose of warfarin is insufficient to protect against thromboembolic events.

3.2.1.) Mechanism of Action

Anticoagulant warfarin inhibits the synthesis of vitamin K dependent coagulation factors (e.g., Factors II, VII, IX and X) and also the anticoagulant proteins C and S.

Table 3.1. Half-life of blood factors and proteins involved in coagulation.

Coagulation Factors and Proteins	Half-Life (hour)
Factor II	60
Factor VII	4-6
Factor IX	24
Factor X	48-72
Protein C	8
Protein S	30

Warfarin suppresses the process of coagulation by decreasing the production and activities of the coagulation factors by inhibition of the regeneration of vitamin K1 epoxide. Vitamin K is an essential cofactor for the gamma-carboxylation of the coagulation factors.

3.2.2.) Pharmacokinetics and Pharmacodynamics

Pharmacodynamics (PD)

Based on the current evidence, warfarin inhibits C1 subunit of the vitamin K epoxide reductase (VKORC1) enzyme resulting in reduction in synthesis of

clotting factors as well as vitamin K1 epoxide. This effect is proportional to the VKORC1 genotypic profile of patients and also the dosage of the medication. "Therapeutic doses of warfarin decrease the total amount of the active form of each vitamin K dependent clotting factor made by the liver by approximately 30% to 50%. An anticoagulation effect generally occurs within 24 hours after warfarin administration. However, peak anticoagulant effect may be delayed 72 to 96 hours. The duration of action of a single dose of racemic warfarin is 2 to 5 days. The effects of warfarin may become more pronounced as effects of daily maintenance doses overlap. This is consistent with the half-lives of the affected vitamin K-dependent clotting factors and anticoagulation proteins: Factor II - 60 hours, VII - 4 to 6 hours, IX - 24 hours, X - 48 to 72 hours, and proteins C and S are approximately 8 hours and 30 hours, respectively" (Coumadin, 2011).

Pharmacokinetics (PK)

Warfarin is a racemic mixture of the R- and S-enantiomers. The clearance of S-enantiomer in the body is much quicker than R-enantiomer. It is also a more active anticoagulant component compared to R- enantiomer (i.e., 2 to 5 times more active).

Absorption

Warfarin's maximum blood concentration is usually attained within 4 hours after oral administration. Warfarin is completely absorbed after oral administration.

Distribution

Almost all warfarin binds to plasma proteins. According to the warfarin's label (Coumadin, 2011), "using a one compartment model, and assuming complete

bioavailability, estimates of the volumes of distribution of R- and S-warfarin are similar to each other and to that of the racemate."

Metabolism

According to the warfarin's label (Coumadin, 2011), "the elimination of warfarin is almost entirely by metabolism. Warfarin is stereoselectively metabolized by hepatic microsomal enzymes (cytochrome P-450) to inactive hydroxylated metabolites (predominant route) and by reductases to reduced metabolites (warfarin alcohols). The warfarin alcohols have minimal anticoagulant activity. The metabolites are principally excreted into the urine; and to a lesser extent into the bile. The metabolites of warfarin that have been identified include dehydrowarfarin, two diastereoisomer alcohols, 4'-, 6-, 7-, 8- and 10-hydroxywarfarin. The cytochrome P-450 isozymes involved in the metabolism of warfarin include 2C9, 2C19, 2C8, 2C18, 1A2, and 3A4. 2C9 is likely to be the principal form of human liver P-450 which modulates the in vivo anticoagulant activity of warfarin. The S-enantiomer of warfarin is mainly metabolized to 7-hydroxywarfarin by CYP2C9, a polymorphic enzyme."

CYP2C9 variant alleles have significant effects on the metabolism of warfarin. The single or multiple variant alleles of this gene (e.g., *1/*2, *1/*3, *2/*2, *2/*3, *3/*3) decrease metabolism of warfarin through lower CYP2C9 enzymatic 7-hydroxylation of S-warfarin and results in decreased S-warfarin clearance (Table 3.2, Yasar, 1999).

Excretion

The half-life of R-warfarin ranges from 37 to 89 hours, while that of S-warfarin ranges from 21 to 43 hours. More than 90% of the received warfarin is detectable in urine and is excreted through urine in the form of metabolites.

Table 3.2. Relationship between S-Warfarin Clearance and CYP2C9
Genotype in Caucasians Patients (Yasar, 1999).

CYP2C9 Genotype	Number of Study Subjects	S-Warfarin Clearance/Lean Body Weight (mL/min/kg) Mean(SD)
*1/*1	118	0.065(0.025)
*1/*2 or *1/*3	59	0.041(0.021)
*2/*2, *2/*2, or *3/*3	11	0.020(0.011)
Total	188	

Elderly

According to different dosing guidelines (e.g., AGSCP Guidelines, 2000), the
use of warfarin in older people requires special consideration as their PT/INR
response to the anticoagulant effects of warfarin is greater than expected. It is
unknown why they are more sensitive to warfarin. Therefore, as patient age
increases, a lower dose of warfarin is usually required to reach the therapeutic
INR level.

Asians

Asian patients may require lower initiation and maintenance doses of warfarin.
In a number of studies, it has been observed that Chinese patients require
lower warfarin dosages to achieve an INR of 2-2.5 (Veenstra, 2005). Some
studies also have shown that the most important determinant of warfarin
dosage in Chinese patients is age (e.g., Veenstra, 2005).

Renal Dysfunction

Given that renal clearance has minor effect body's response to warfarin, there
is no need to make dose adjustment if a patient had renal failure.

Hepatic Dysfunction

Liver has an important role in the metabolism of warfarin. Accordingly, any
hepatic dysfunction, which usually results in impairment in the synthesis of

clotting factors as well as decreased warfarin metabolism, can enhance body response to warfarin.

3.2.3.) Pharmacogenomics

In 2005, Sanderson team reported the results of a meta-analysis of 9 studies (2775 patients: 99% Caucasian) performed to investigate the CYP2C9 gene variants-associated clinical outcomes in warfarin-treated patients (Sanderson, 2005). In this meta-analysis, some of the qualified included studies assessed bleeding risks and the rest of the studies assessed daily dose requirements. The analysis suggested a higher risk of bleeding risk and lower mean daily dose of warfarin for patients with either the CYP2C9*2 or CYP2C9*3 alleles. Patients with CYP2C9*2 and CYP2C9*3 alleles have lower mean daily warfarin doses and a greater risk of bleeding, 17% and 37% less than the mean daily dose for patients homozygous for the CYP2C9*1, respectively.

In a prospective study of 219 Swedish patients under warfarin treatment who were stratified by CYP2C9 genotype, Lindh et al. discovered that the risk of overanticoagulation (i.e., achieving INR >3) during the first 2 weeks of warfarin therapy was two times higher for the patients with the polymorphic variant alleles (CYP2C9*2 and CYP2C9*3) compared to patients with wild-type CYP2C9 (Lindh, 2005).

Many studies have found that patients with some of the VKORC1 gene's single nucleotide polymorphisms (e.g., -1639G>A, rs9923231) require lower initial doses of warfarin. Gene VKORC1 regulates the synthesis of the vitamin K–epoxide reductase (VKOR) protein which is the target enzyme of warfarin (Rost, 2004; Li, 2004). The -1639G>A polymorphism alters the biding sites of the VKOR which results in a reduction in protein expression (Rieder, 2005;

Yuan 2005). A large number studies have discovered that the variants of VKORC1 and CYP2C9 genes are individually responsible for 35% to 50% of the variable dose response to warfarin. For example, Wadelius team reported an association between VKORC1 gene variations and lower required dose of warfarin. In this study with a study cohort of 201 white patients under warfarin, they attributed about 30% and 40% of the variance in warfarin dose to variations in the VKORC1 gene and variations in VKORC1 and CYP2C9 genes combined, respectively (Wadelius, 2005). Several multivariate analyses have shown that the addition of patient characteristics, such as age, gender, height, weight and other medications, to CYP2C9 and VKORC1 accounts for approximately 50-60% of warfarin dose variability.

3.2.4.) Accounting for Adverse Reactions

Oral anticoagulants, most commonly warfarin, reduce risk to thrombosis and treat conditions that might lead to stroke, pulmonary embolism, deep vein thrombosis or other blood clotting related disease. The impact and value of anticoagulation medication in the U.S. is dramatic. For example, stroke is the third leading cause of death in the U.S. with over 140,000 deaths annually. The majority of stroke incidences are due to ischemia (87%) or transient ischemic attack (TIA, ~5-10%) and are typically managed by the use of anticoagulation agents such as warfarin, dabigatran, and clopidogrel. Whatever the patient's disease or condition leading to a prescription of an anticoagulation agent, selecting the best combination of drug and treatment protocol is complicated by the individual differences in anticoagulation medication response (e.g. >20-fold difference for warfarin) due to genetics, physiology, and compliance. Consequently, given this characteristic, any

small increase or decrease in medication dosage might increase the risks of bleeding or thromboembolic events, respectively (Ansell 2008). In practice, providers use a combination of experience, scientific evidence and clinical trial results to develop anticoagulation "best practice" treatment plans designed to roughly minimize the patient-to-patient response variability and risks across the provider's patient population. However, the high degree of patient heterogeneity causes variations in individual patient response to these "best practice" drug-protocol approaches. Another factor that makes the management of warfarin more complex is its interactions with many other medications and foods. Therefore, warfarin is among the medications with a high rate of associated adverse reactions (Ansell, 2008).

According to the 2010 FDA-approved warfarin (Coumadin) product label (Coumadin, 2011), there are a wide range of adverse reactions associated with warfarin:

- Fatal or nonfatal bleeding from any tissue or organ. Depending on the severity and location of the bleeding, the complications may present as paralysis; paresthesia; headache, chest, abdomen, joint, muscle or other pain; dizziness; shortness of breath, difficult breathing or swallowing; unexplained swelling; weakness; hypotension; or unexplained shock. One important point with regard to bleeding during anticoagulation therapy is that it does not always correlate with PT/INR. In this case, the bleeding might have been resulted from other disorders such as tumors and ulcers.

- Necrosis of skin and other tissues.

- Infrequent or rare adverse reactions: hypersensitivity/allergic reactions, including anaphylactic reactions, systemic cholesterol microembolization,

purple toes syndrome, hepatitis, cholestatic hepatic injury, jaundice, elevated liver enzymes, hypotension, vasculitis, edema, anemia, pallor, fever, rash, dermatitis, including bullous eruptions, urticaria, angina syndrome, chest pain, abdominal pain including cramping, flatulence/bloating, fatigue, lethargy, malaise, asthenia, nausea, vomiting, diarrhea, pain, headache, dizziness, loss of consciousness, syncope, coma, taste perversion, pruritus, alopecia, cold intolerance, paresthesia including feeling cold and chills, tracheal or tracheobronchial calcification, and priapism.

3.2.5.) Clinical Protocols

The narrow therapeutic range and wide interindividual variability in warfarin therapeutic dose (such as 4.5–77 mg/week (Wadelius, 2004)) make anticoagulation management challenging and anticoagulation response unpredictable. Current clinical best practice relies primarily on empirical dosing. Accordingly, most patients usually start taking a fixed dose each day (such as 5 mg/day) during the "initiation phase" of warfarin on the basis of population averages, regardless of clinical and genetic factors (Garcia, 2005). Then, based on the INR results, the dose is titrated. This empirical clinical practice approach requires frequent changes in the dose of warfarin in response to out-of-therapeutic range INRs and to avoid adverse effects and maintain therapeutic efficacy. To address this challenge, mainly before the completion of the Human Genome Project, a number of dosing algorithms that included clinical variables were developed. Since the successful completion of the project and in the light of discoveries of polymorphisms in cytochrome P450 2C9 (CYP2C9) and vitamin K epoxide reductase complex 1 (VKORC1) which jointly account for about 40-50% of the inter-individual variability in dose

requirements, a large number of pharmacogenetic-based dosing algorithms have been also developed. However, the potential benefit of these dosing algorithms in terms of their safety and clinical utility has not been adequately investigated in randomized settings.

3.2.6.) Dosage and Administration

The dosage and administration of warfarin must be individualized for each patient. The warfarin dose management is an iterative process which starts with administering an initial dose and then is followed with dose adjustments based on the patient's INR response to the medication (Figure 3.9).
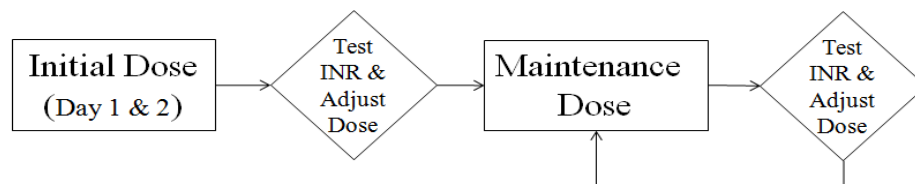
Figure 3.9. The iterative process of dosing and adjusting warfarin dose until achieving maintenance dose and INR in therapeutic range.

As discussed earlier, many factors cause warfarin dose variability such as clinical factors including age, race, body weight, sex, concomitant medications, comorbidities and diet and genetic factors including CYP2C9 and VKORC1 genotypes. To initiate a warfarin dose, there are different options such as:

- According to FDA, if the patient's genotypes are known, we can use the following table (Table 3.3) to select the initial dose. If the genotypes are not available, the initial dose of warfarin is usually 2 to 5 mg/day. This dose should be modified based on consideration of patient-specific clinical factors.

Table 3.3. Recommended daily warfarin doses (mg/day) to achieve a therapeutic INR based on CYP2C9 and VKORC1 genotype using the warfarin label approved by the US Food and Drug Administration (Johnson, 2011).

VKORC1 Genotype	CYP2C9 Genotype					
(-1639G>A, rs9923231)	*1/*1	*1/*2	*1/*3	*2/*2	*2/*3	*3/*3
GG	5-7	5-7	3-4	3-4	3-4	0.5-2
GA	5-7	3-4	3-4	3-4	0.5-2	0.5-2
AA	3-4	3-4	0.5-2	0.5-2	0.5-2	0.5-2

- Use any available clinical or PG-based dosing algorithms to calculate appropriate initial dose or

- Use the standard fixed-dose practice (5mg/day, Loading Dose: 10 mg/day). After initiating warfarin, the subsequent dosage adjustments must be made based on INR results until achieving a maintenance dose and an INR in therapeutic range.

3.2.7.) PK/PD Modeling and Simulation (INR Prediction)

In order to model the pharmacodynamics (PD) and pharmacokinetics (PK) of a medication, the relationship between drug dose, plasma concentration, biophase concentration (pharmacokinetics), and drug effect or side-effects (pharmacodynamics) is characterized, and relevant patient covariates are included in the model (Figure 3.10). Modeling of pharmacokinetic and
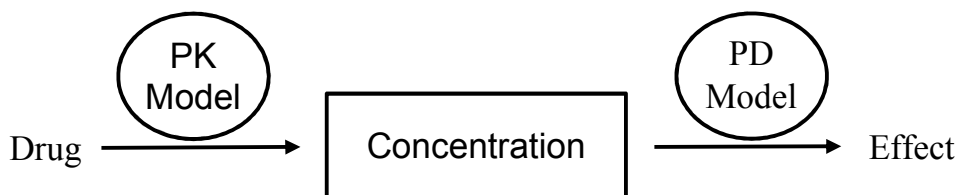
Figure 3.10. A schema depicting the general relationship of pharmacokinetic (PK) and pharmacodynamic (PD) models used in clinical pharmacology.

pharmacodynamic can be used to make predictions about the "temporal profile of the drug concentration" and "its effect" which ultimately helps select

appropriate dosing protocol and "optimal delivery profile" (Yan, 2013).

A number of studies have been undertaken to model PK and PD of warfarin.
The aim of most of these studies have been to develop population models to
describe the PKs of both S- and R-warfarin and the PK-PD relationship
between warfarin exposure/concentrations and anticoagulant response (i.e.,
INR), including identification of important predictors for a priori dose
individualization of warfarin.

In the next following section before reviewing one of the warfarin's PK/PD
models, We are going to briefly go over some fundamental concepts of PK
and PD.

Pharmacokinetic Models

Pharmacokinetics (PK) studies the relationship between drug concentration
and its effect on the body. PK principles are often used to reduce toxicity and
improve efficacy during patient care. The suite of tools used in PK studies are
mathematical models, used to quantify the processes of drug Absorption,
Distribution, Metabolism, and Excretion (ADME) in the body. Rates of
reactions and compartmental architecture are two important properties of PK
models that control how these ADME processes are modeled.

Rates of Reaction

The ADME processes can be modeled as either zero-order or first-order
reactions. The order of the reaction is the rate of change of a variable over
time. Consider an example where drug A is modeled. In a zero-order reaction,
the rate of change of drug A would be constant (see equation A), where k^* is a
zero-order rate constant. Here the rate of change of A is independent of the
concentration of A. In a first-order reaction, the rate of change of drug A is

proportional to A (see equation B), where *k* is a first-order rate constant (Dhillon, 2006).

$$\frac{d_A}{d_t} = -k^*$$ (A)

$$\frac{d_A}{d_t} = -kA$$ (B)

The drug's rate of reaction has important clinical implications. Drugs that have first-order rate of elimination do not accumulate within the body because as concentration increases so does the elimination rate. In contrast, drugs with zero-order rates of elimination will accumulate with continued administration. However, a first-order reaction can be altered to appear as a zero-order reaction in such situations as overdosing. At clinical dosages though, most drugs are first-order reactions, with a few zero-order examples such as phenytoin and high-dose salicylates.

Compartmental Models

The compartmental architecture of a PK model determines the fate of a drug after it enters the body. Although the compartments are hypothetical structures to model the body, they do have underlying biological reasoning to describe how a drug is processed. In reality, drug concentration and kinetics will vary with the type of tissue (e.g. brain versus muscle). Therefore, to accommodate different modeling scenarios there are several types of compartmental architectures.

One-compartment Model

The one-compartment model represents the body with one compartment and makes the simple assumption that once the drug is introduced it is

Figure 3.11. One-compartment model. k_a = absorption rate constant, k = elimination rate constant. Adopted from (Dhillon, 2006).

instantaneously and homogeneously distributed throughout the body (Figure 3.11, Dhillon, 2006). The drug concentration is monophasic and decreases exponentially with time, or linearly if the log of drug concentration is taken (Figure 3.12, Dhillon, 2006).
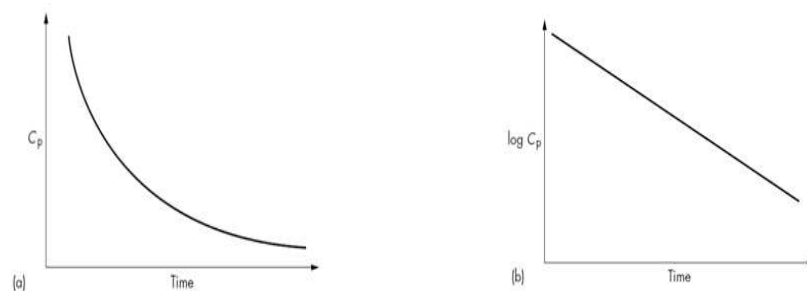
Figure 3.12. (a) Plasma concentration (C_p) versus time. (b) Time profile of a one-compartment model showing log C_p versus time. Adopted from (Dhillon, 2006).

Two-compartment Model

The two-compartment models decompose the body into two tissues types: the central component represents highly perfuse tissue (e.g. heart, kidneys and lung) and the peripheral component represents tissues that are less perfuse (e.g. skin, muscle and fat). The two-compartment model allows more complex drug concentration dynamics than the one-compartment model. Upon administration, the drug's concentration has a two phases: it is initially highly concentrated in the central component but rapidly declines as it distributes to the peripheral component. After reaching equilibrium between the two

compartments, the drug then declines more slowly as it is eliminated from the central compartment. Hence, the drug concentration over time shows a biphasic response (Figure 3.13, Dhillon, 2006).
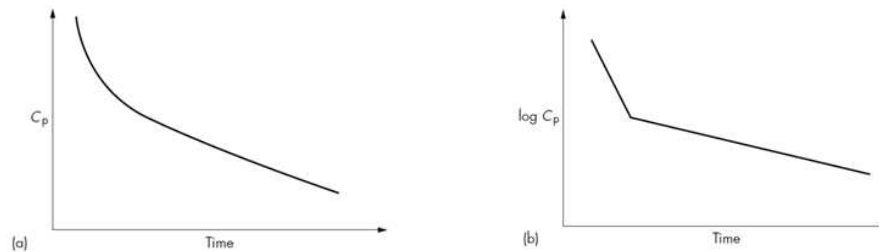
Figure 3.13. (a) Plasma concentration versus time profile of a drug showing a two compartment model. (b) Time profile of a two-compartment model showing log C_p versus time. Adopted from (Dhillon, 2006).

Multi-compartment Model

The multi-compartment models allow even more complex drug concentration dynamics. With higher number of compartments there will be more phases in the drug concentration over time (Figure 3.14).
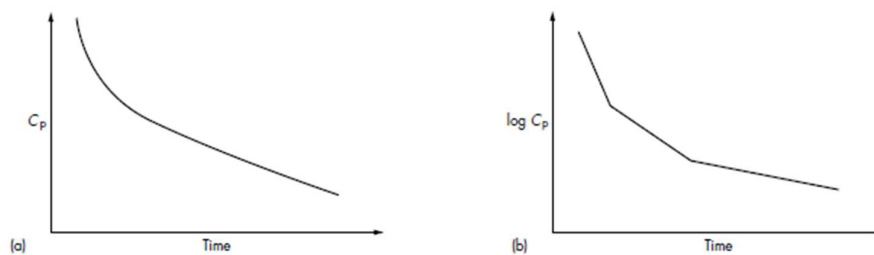
Figure 3.14. (a) Plasma concentration versus time profile of a drug showing multi-compartment model. (b) Time profile of a multi-compartment model showing log C_p versus time. Adopted from (Dhillon, 2006).

Pharmacodynamics

Pharmacodynamics studies the drug's effect on the body by describing how the drug affects local physiological and biochemical processes in the body. At the core of the pharmacodynamics approach is the reaction equation (see

equation 1) describing the relationship between the drug (D), receptor (R), drug-receptor complex (DR), and the effect (E), where k_{on} and k_{off} are the rate of drug and receptor association and the rate the complex dissociation, respectively. Of course more complex kinetics is possible, but even with such a simple expression some very useful clinical information can be gleaned. The responsibility of researchers and clinicians is then to tailor the mathematical model describing the reaction equation to their pharmacodynamics scenario (Dhillon, 2006).

The expression in equation C has several clinically relevant properties.

$$D + R \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} DR \overset{k_2}{\rightarrow} E \qquad (C)$$
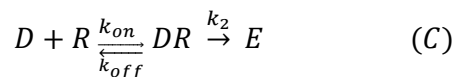
The receptor dissociation rate constant (k_D) describes the equilibrium rate between the rate of k_{on} and k_{off} (equation D). This principal demonstrates, for example, that if [D] is very high then the receptors are saturated, and no significant increase in [E] can take place. The idea that there can be a limit to the amount of effect is quantified with E_{max}, the maximal effect for all drugs. There is another, related value denoted EC_{50}, which quantifies the concentration at which E is 50% of E_{max}. These constants are important for determining the relationship between drug concentration and effect, and how pharmacodynamics models are put into use (Dhillon, 2006).

$$\frac{K_{off}}{K_{on}} = K_d = \frac{[D][R]}{[DR]} \qquad (D)$$

Different pharmacodynamics scenarios dictate the mathematical form describing equation C. Effect can be sigmoidal or linearly related to drug concentration depending on modeling assumptions. The most general equation is shown in equation E, a sigmoidal relationship (Dhillon, 2006).

$$E = \frac{E_{max} \cdot C^n}{(EC_{50}^n + C^n)} \qquad (E)$$

C represents the drug concentration, and n is the "Hill" coefficient, determining the "steepness" of the concentration-effect relationship. Very high values of n (n>5) can make the relationship so steep as to effectively make the drug have a binary effect (i.e. the effect is present or not). Equation F represents the case of n=1 (Dhillon, 2006).

$$E = \frac{E_{max} \cdot C}{(EC_{50} + C)} \qquad (F)$$

Equation G simplifies equation 3 to a linear relationship. This may be useful in situations where C is much less than EC_{50} and the range of clinical dosage is very narrow (Dhillon, 2006).

$$E \approx \frac{E_{max}}{EC_{50}} \cdot C = Slope \cdot C \qquad (G)$$

The log of equation G can be taken to yield equation H (Dhillon, 2006).

$$E = Slope \cdot \log(C) \qquad (H)$$

A model describing a log-linear relationship can be useful in situations where there is high intrinsic variability, but the drawbacks of such models include the inability to represent the case where C is zero such as in placebo studies. The fundamental concentration-effect relationships are outlined in Figure 3.15.

Continuous PD models

PD models can be either continuous or categorical, although the focus here is on continuous models. The type of model used depends on the nature of the data, such as whether the data is continuous (e.g., blood pressure and weight), categorical (e.g., grade of adverse event), and the frequency of measurement. The different concentration-effect relationships discussed
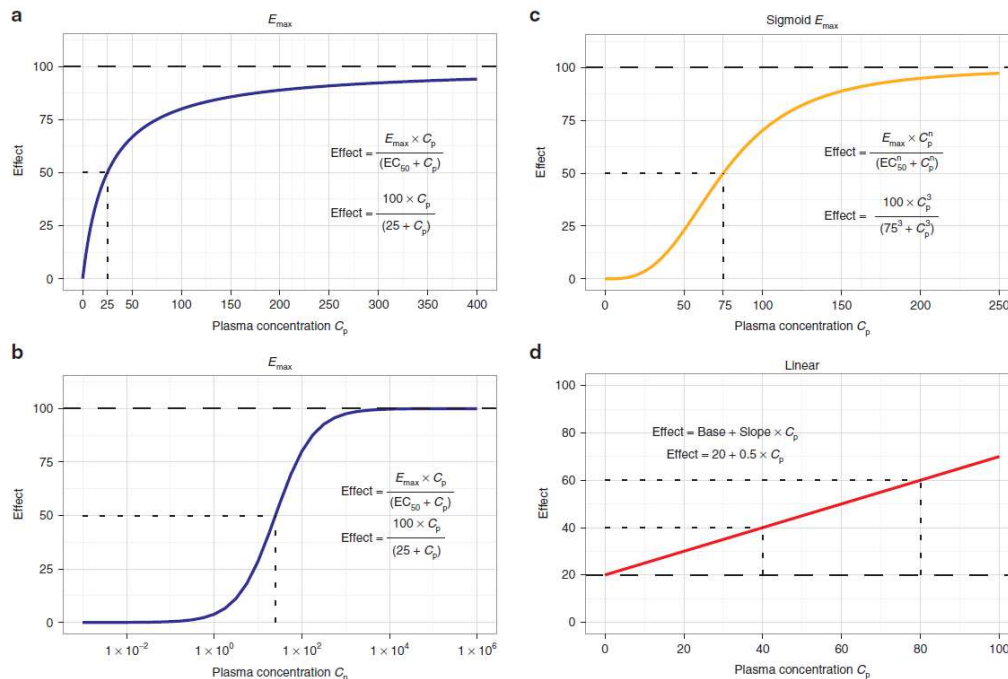
Figure 3.15. Fundamental concentration-effect relationships. (a) The E~max~ concentration-effect relationship that arises from receptor theory for the binding of a single drug to a single receptor. Fifty percent of the maximum effect is achieved at the EC₅₀ concentration. (b) The same relationship over a much wider log-concentration scale. (c) The sigmoid E~max~ concentration-effect relationship that arises from receptor theory when there is allosteric inhibition or simulation of binding. The "Hill factor" n controls the steepness of the middle part of the curve. (d) A linear concentration-effect relationship that is a semiempirical but sometimes useful substitute for an E~max~ relationship when the range of concentration is relatively small and the drug effect is well below E~max~. Adopted from (Upton, 2014).

above provide crucial flexibility for different pharmacodynamics scenarios, but more building blocks are needed to accommodate the diverse physiological context the concentration-effect dynamics take place. Figure 3.16 outlines five common types of continuous PD models in use.
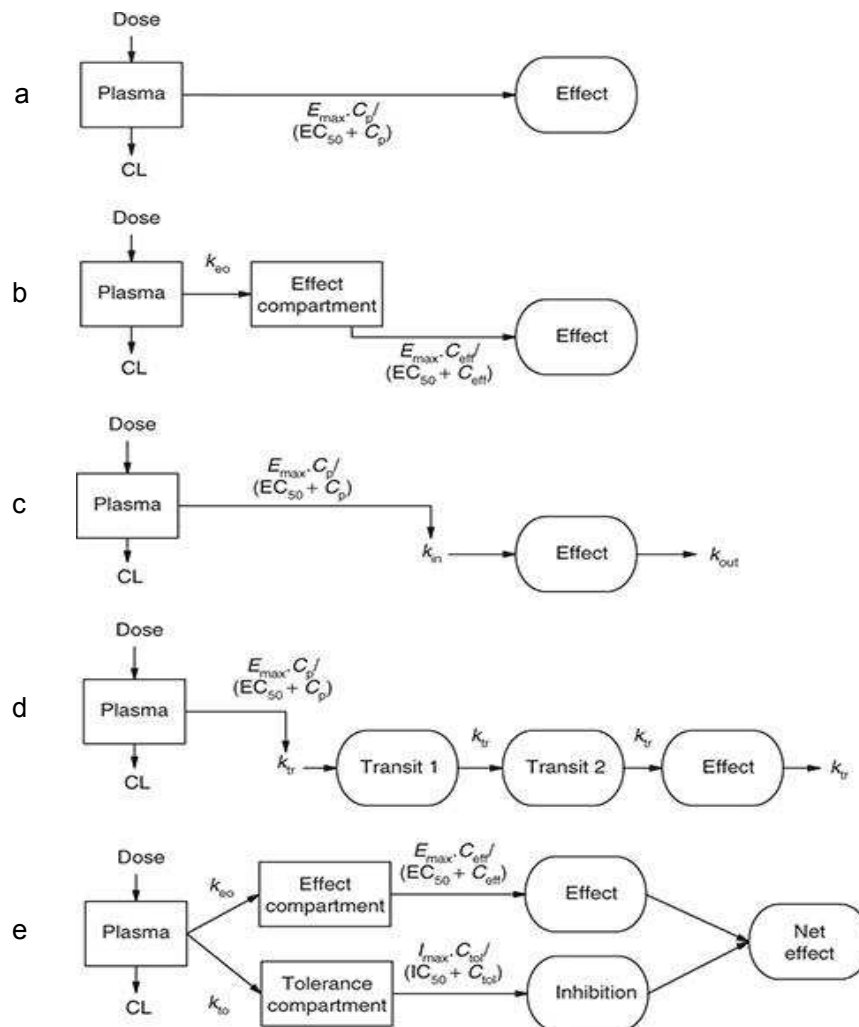
Figure 3.16. Representative continuous pharmacodynamic models. (a) A direct response model where effect is driven by the plasma drug concentration. (b) An effect compartment model where effect is driven by the effect compartment drug concentration, which is delayed relative to the plasma concentration by a first-order rate constant k_{e0}. (c) A turnover model where drug effect is a balance between an apparent production rate (k_{in}) and an apparent removal rate (k_{out}). Drug affects the net effect by altering kin or (k_{out}). (d) A transit compartment model, where the drug effect is at the end of chain of processes and drug action is on the first process. (e) A tolerance compartment model, where the drug effect is described by an effect compartment and the development of tolerance is described by a slower inhibitory compartment that reduces the net drug effect with time. Adopted from (Upton, 2014).

Hamberg's PK/PD Model

The aim of the Hamberg's study (Hamberg, 2007) was to characterize the

relationship between warfarin concentrations and international normalized

ratio (INR) response and to identify predictors important for dose

individualization. S- and R-warfarin concentrations, INR, and CYP2C9 and

VKORC1 genotypes from 150 patients were used to develop a population

pharmacokinetic/pharmacodynamic model in NONMEM (Beal, 1994). The

anticoagulant response was best described by an inhibitory EMAX model,

with S-warfarin concentration as the only exposure predictor for response.

Delay between exposure and response was accounted for by a transit

compartment model with two parallel transit compartment chains (Figure

3.17).
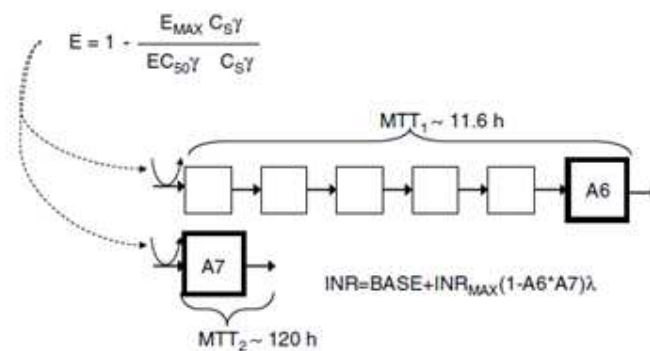
Figure 3.17. A two-chain transit compartment PD model used by Hamberg et al. to describe the INR response to warfarin therapy. Mean Transit Times (MTTs), Apparent clearance of S-warfarin [CLs], EC50 (concentration resulting in 50% of Emax), λ is the sigmoidisity factor, describing the steepness of the concentration–effect relationship, A1–7 indicate compartment amounts, A1–6 are the compartment amounts in the "long" transit chain, whereas A7 is the compartment amount in the "short" transit chain (Hamberg, 2007).

They also found a two-compartment PK model with first-order input and first

order elimination to appropriately characterize the disposition of S-warfarin

(Figure 3.18).
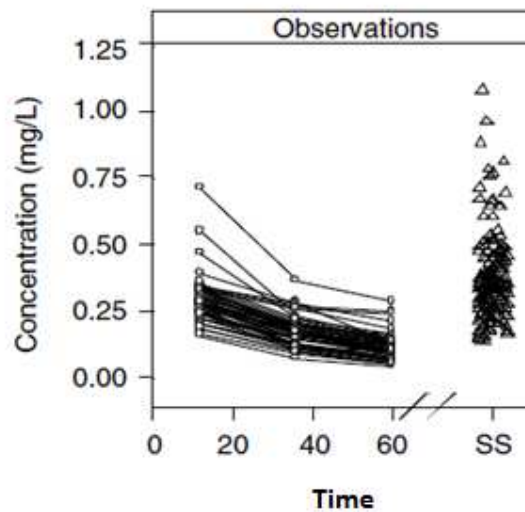
Figure 3.18. Observed time profiles of S-warfarin
concentrations following administration of a single 10 mg
warfarin dose and at steady state (Hamberg, 2007).

In their study, CYP2C9 genotype and age were identified as predictors for S-warfarin clearance, and VKORC1 genotype as a predictor for warfarin sensitivity. Accordingly they have modeled the ith individual's jth observed INR value (INR_{ij}) according to:

$$INR_{ij}=BASE_i+INR_{MAX}(1-A6*A7)^\lambda$$

$BASE_i$ is the ith individuals observed baseline INR value whose predictive covariates are Age, CYP2C9 and VKORC1, INR_{MAX} is the maximum INR increase from baseline, λ is the sigmoidisity factor, describing the steepness of the concentration - effect relationship, A1–7 indicate compartment amounts, A1–6 are the compartment amounts in the "long" transit chain, whereas A7 is the compartment amount in the "short" transit chain (Hamberg, 2007). According to Hamberg, predicted INR curves show significant steady-state differences across patients with different covariates. They indicated that

it was not possible to anticipate these differences based on just early INR assessments.

The following table (Table 3.4) depicts warfarin doses predicted using Hamberg's model for 54 individuals with different set of predictive covariates of CYP2C9 and VKORC1 genotypes for three age groups of 50, 70, and 90 years old. The desired target INR in this effort was 2.5. As an example, the predictions show the significant difference (20 fold) between *a priori* doses (9.08 mg/day vs. 0.47 mg/day) for two individual with different combination of predictive covariates ("A 50-year-old patient with CYP2C9 *1/*1 and VKORC1 GG genotypes" vs. "A 90-year-old patient with CYP2C9 *3/*3 and VKORC1 AA genotypes").

Table 3.4. Predicted daily warfarin dose with target INR of 2.5 in three groups of patients with different set of predictive covariates (i.e., Age, CYP2C9 and VKORC1 genotypes) (Hamberg, 2007)

	VKORC1(-1639)								
	G/G			G/A			A/A		
CYP2C9	50 yrs	70 yrs	90 yrs	50 yrs	70 yrs	90 yrs	50 yrs	70 yrs	90 yrs
*1/*1	9.08	7.72	6.36	5.94	5.06	4.16	4.32	3.68	3.04
*1/*2	6.22	5.3	4.4	4.1	3.5	2.9	2.98	2.54	2.1
*1/*3	5.04	4.3	3.58	3.3	2.82	2.34	2.4	2.06	1.71
*2/*2	2.54	2.16	1.77	1.66	1.41	1.16	1.21	1.03	0.84
*2/*3	2.82	2.4	1.97	1.85	1.57	1.29	1.35	1.14	0.94
*3/*3	1.38	1.18	1	0.9	0.77	0.65	0.66	0.56	0.47

The authors indicated that the influence of warfarin therapy on INR response of the R warfarin was not statistically significant. Therefore, the model only considered the PK/PD effects of S-warfarin. They concluded that it is

important to account for CYP2C9 and VKORC1 genotypes and age to improve *a priori* and *a posteriori* individualization of warfarin therapy.

PK-PD modeling for INR response in our *in silico* framework

We have adapted the PK-PD model from Hamberg (Hamberg, 2007) to predict the INR response for each clinical avatar at a given warfarin dose. In order the make the INR model work within our framework, we ported the code from NONMEM to R. According to the model, we only considered the PK-PD effects of S-warfarin. We modeled the PK effects using a two-compartment model with first order input and first order elimination and the PD effects using a two-chain transit compartment model. Due to limitations on the original model it has been necessary to make assumptions about the covariance of the variables because the complete covariance matrix was not provided. We have used a random log normal distribution to estimate the variability of the clearance rate, the volume in the central compartment, and the volume in the peripheral compartment and restricted the range to be within physiological ranges.

To model the accumulation of warfarin concentration over time (assuming daily doses), we have used the principle of superposition. Superpositioning does not require assumptions regarding a PK model or absorption kinetics, but instead assumes each dose of the drug acts independently and that the rate and extent of absorption and average systemic clearance are the same for each dosing interval and that linear PK apply (Gibaldi, 1982). We create a table of warfarin concentrations over time and summed across the rows at 24-hour time intervals to predict the amount of warfarin remaining in the system.

3.3.) Design of Comparative Effectiveness and Clinical Trial Studies

A large number of the definitions have been offered for Comparative Effectiveness Research (CER). According to the Center for Clinical and Translational Science of Ohio State University, CER is "a field of study that utilizes data generation (new studies) and synthesis (comparisons of existing studies) to provide evidence identifying best practices and policies related to improving health care." CER studies are generally conducted after randomized controlled trials (RCTs) (CER Resources, 2014). The focus of most of RCTs has been to study the efficacy of treatments under ideal conditions. On the other side, CER studies mostly focus on effectiveness by comparing one or more treatments, tools, procedures or medications to determine what works best for which individuals or patient populations under real world situations. CER studies have gained significant attention in clinical medicine in last few years since the Patient Protection and Affordable Care Act (PPACA, 2010) established Patient-Centered Comparative Effectiveness Research (PC-CER) as a US national medical research priority. Although this law has provided tremendous resources for CER studies to improve the evidence base that supports the use of genomic information to improve healthcare, however, efforts to translate critical genomic discoveries require prohibitively expensive clinical trial and clinical study validation which are severely hindered by regulatory, technical and validation barriers not easily conducted using current clinical-research or clinical enterprise environments. According to Khoury (2007), the reality is that "a small proportion of human genomics research has progressed from gene discovery to an evidence-

based health application that has been effectively integrated into practice and has demonstrated health impact." Based on Tunis (2003), to address this situation and make the clinical trials' results more efficient and generalizable, there is a need for CER studies in which we (1) select clinically relevant alternative interventions to compare, (2) include a diverse population of study participants, (3) recruit participants from heterogeneous practice settings, (4) collect data on a broad range of health outcomes, and (5) pay special attention to the study time frame.

One "poster" example of this translational complexity and need for a new approach to predicting a population wide treatment plan is the case of "optimal" warfarin dosing (~30 million US prescriptions/year) which demonstrates highly variable individual risks to serious under and over dosing adverse responses (thrombotic or bleeding event) and rapidly increasing health care costs associated with warfarin complications estimated at over $1 billion/year (McWilliam, 2006). The clinical potential of genetics has driven significant scientific and clinical efforts to study the warfarin dosing question. Key scientific findings have demonstrated significant relationship between genotypes of VKORC1 and CYP2C9 and the metabolism of warfarin (e.g. Rieder, 2005). Consequently, the genotype-dosing relationship of warfarin as one of the emerging collection of pharmacogenomic results has led to a large number of clinical trials. As a result, various warfarin PG-based dosing algorithms and protocols have been offered and many outcome metrics (e.g, INR and TTR) have been tested in different related studies. However, given that anticoagulation therapy is usually a long term treatment and the fact that clinical utility of the PG-based warfarin algorithms has been tested only in

small study populations and highly constrained conditions (Anderson, 2007; Caraco, 2008; Burmester, 2011; Wang 2012), conducting longitudinal CER studies including multiple dosing protocols are still necessary to evaluate the ultimate impact of the available warfarin dosing algorithms on practice and different populations (i.e., external validity). These studies could address questions such as how and under which conditions PG-based dosing algorithms could perform well in certain populations for whom PG-based algorithms have shown less effective performance compared to other populations (Schelleman, 2008; Gage, 2008; Cavallari, 2012).

3.3.1.) Longitudinal Studies

Long-term warfarin anticoagulation is commonly used to prevent thromboembolism in patients with medical conditions such as atrial fibrillation and venous thromboembolism (e.g., deep vein thrombosis (DVT) or pulmonary embolism), or when treating patients with mechanical heart valves (MHVs) (Daniels, 2005). Long-term use of warfarin along with its narrow therapeutic index and a high risk/benefit ratio necessitate close and long-term monitoring. Through monitoring, different temporal quality measures are examined: the patterns of warfarin use in terms of discontinuations and interruptions, quality of anticoagulation therapy using primary anticoagulation outcome metrics (e.g., INR and TTR) as well as the relationship of these patterns with subsequent stroke and bleeding events as secondary and clinical outcomes. To study these patterns, in addition to prospective clinical trials, the secondary use of EMR data provides a great opportunity to design and conduct observational longitudinal studies or retrospective longitudinal cohort studies. Using treatment data from longitudinal studies, we can

estimate quantitative and temporal parameters associated with the quality of anticoagulation therapy. It helps to individualize doses and consistent and efficient dose adjustment practices using PG-based or non PG-based dosing protocols. This kind of studies could be conducted not only to study the effectiveness of different PG-based dosing algorithms for initiating warfarin doses but also they could be done to address questions such as how long genotype remains a significant predictor of warfarin dose.

Our *in silico* translational research framework could take advantage of longitudinal EMR data of warfarin patients to conduct the CER studies that could address the effectiveness of different dosing algorithms/protocols over the period or treatment. It is discussed in more detail in chapters 5 and 6.

3.3.2.) INR and TTR Estimation

Percent Time in Therapeutic Range of INR (TTR) for each cohort of patients can be determined using any of the following three methodologies (Table 3.5):

1- The fraction of INR's in range (Loeliger, 1985). The fraction of INR's in range is calculated by taking the number of INR's within target range for all patients divided by the total number of INR's during selected time interval.

2- The Rosendaal linear interpolation method (Rosendaal, 1993). The Rosendaal linear interpolation methodology assumes a linear relationship exists between two INR values and allows one to allocate a specific INR value to each day for each patient. An average time in range for all patients was determined. The Rosendaal method has demonstrated to be valid and reproducible when the level of missing INR values is not high (e.g., ~20%) (Hutten, 1999).

3- The cross-section-of-the-files methodology (Ansell, 2004). The method takes each patient whose INR value is in range at one point in time (the INR value that was closet the midpoint of the selected interval ± 7 days) divided by the total number of INR's done on all patients at that point in time. In other words, it assesses all patients being managed at one point in time by taking the total of those whose INR is in range and dividing it by the total number of patients who had an INR at that point in time.

Table 3.5. Advantages and Disadvantages of Methods to Calculate Time-in-Therapeutic Range (TTR) (Schmitt, 2003).

Methodology	Advantage	Disadvantage
Fraction on INRs	Simple to calculate Requires only one INR value per patient in clinic population Not influenced by extent of INR out-of-range	More frequent testing in unstable patients may bias overall results (will under-estimate TTR of group) Does not take into account days within target range Does not consider individual patients
Cross-section-of-the-files	Simple to calculate Considers individual patients Not influenced by extent of INR out-of-range	Does not take into account actual days within target range Only considers one point in time
Rosendaal linear interpolation	Takes into account actual days in target range Allows one to calculate INR specific incidence rates of adverse events	Calculation more difficult Makes assumptions about INR between actual tests Does not consider individual patients Extreme out-of-range INR values may bias overall results

3.3.3.) Multiple Protocol Studies

Clinical trials have produced a host of treatment protocols for anticlotting medications such as warfarin, where over 40 protocols were published in the past 30 years. In practice, providers use a combination of experience, scientific evidence and clinical trial results to develop anticoagulation "best practice" treatment plans designed to roughly minimize the patient-to-patient response variability and risks across the practice's patient population. However, the high degree of patient heterogeneity (based on factors such as race, age, individual medical data, family history and genetics) causes variations in individual patient response to these "best practice" drug-protocol approaches. The sheer number of treatment options and risk factors preclude the full array of clinical trials required to test all combinations of patients and drug-treatment protocol options. In short, no practical approach to identify the optimal anticoagulation treatment plan exists for large heterogeneous patient populations that accounts for individual risk factors; drug and protocol options; and achieves minimal risk to adverse events such as stroke. Current access to large comprehensive electronic medical records (EMR) covering diverse patient populations, coupled with novel modeling and computational simulations provides an unprecedented opportunity to conduct *in silico* identification and validation of optimal treatment strategies.

To address such a challenge, our *in Silico* WiAD PCOR CER Framework is used to produce large, representative synthetic patient populations that can be used to conduct replicated clinical simulations testing and comparing multiple anticoagulation medication-protocol options.

3.4.) Model and Simulation Requirements

In our adjustable modular anticoagulation therapy simulation framework (Figure 3.19), we use a highly adaptable modeling and software application development paradigm to create models of actual patient populations and then use those models to produce simulated patient populations ("clinical avatar" populations) and conduct simulations to predict the clinical validity and efficacy of genetic tests and algorithms applied to each clinical avatar. The
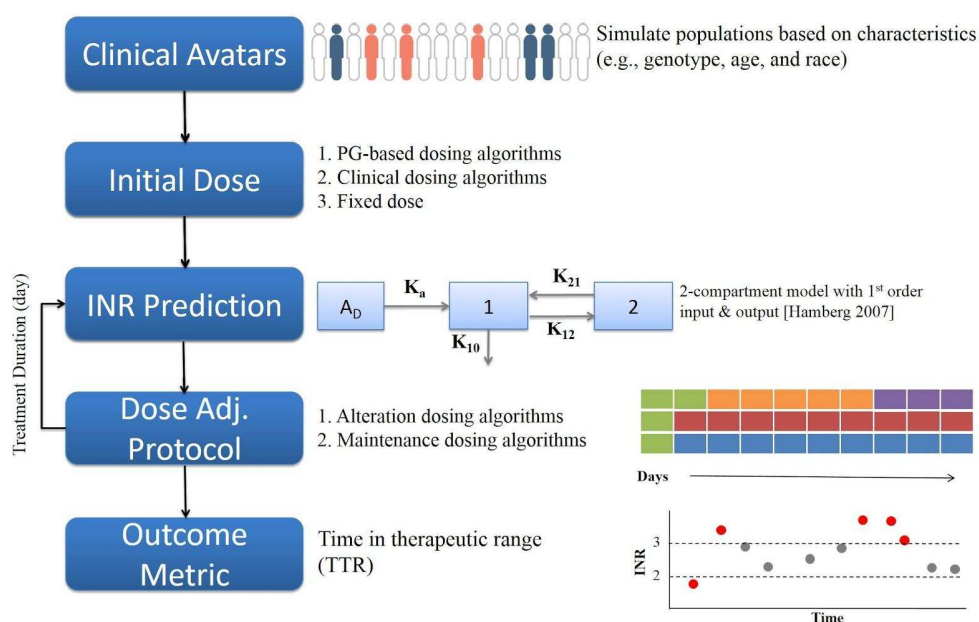
Figure 3.19. Adjustable Modular Anticoagulation Therapy Simulation Framework.

mathematical representation of statistically accurate clinical avatar populations are used to simulate clinical data, warfarin dose (initial and adjusted doses) and INR response over desired period of time for anticoagulation treatment courses using different patient populations and a collection of pharmacogenetic and non-pharmacogenetic dosing algorithms. Different outcome metrics are calculated and examined to determine which

algorithm provides the evidence supporting the best clinical outcome for each patient population.

Next, we will use Aurora Health Care's patient-based electronic medical record to assemble hospital, city, county-wide and regional patient populations. We will use the modeling framework to conduct 90 day simulations on these hospital patient populations and test the predictions of over and under dosed patients against data provided by Aurora Health Care. In addition to the clinical trial simulations, we will also simulate Milwaukee City, County and South East Wisconsin populations and apply standard health disparity and geographical analysis to predict the likelihood of higher incidences of adverse events in geographically and racially diverse sub populations. We will test these predictions against Aurora's stratified patient population to determine accuracy of the population-wide simulations.

Our approach to generate clinical avatars follows the standard applied mathematical modeling approach (explained in detail in section 3.1.2.): Analyze and characterize the data; Formulate a phenomenological model that 'fits' the data; Test the performance of the model against a sub-collection of data; Evaluate the accuracy of the model; and Adjust the model based on the accuracy (or lack thereof) (Blue box in Figure 3.20).
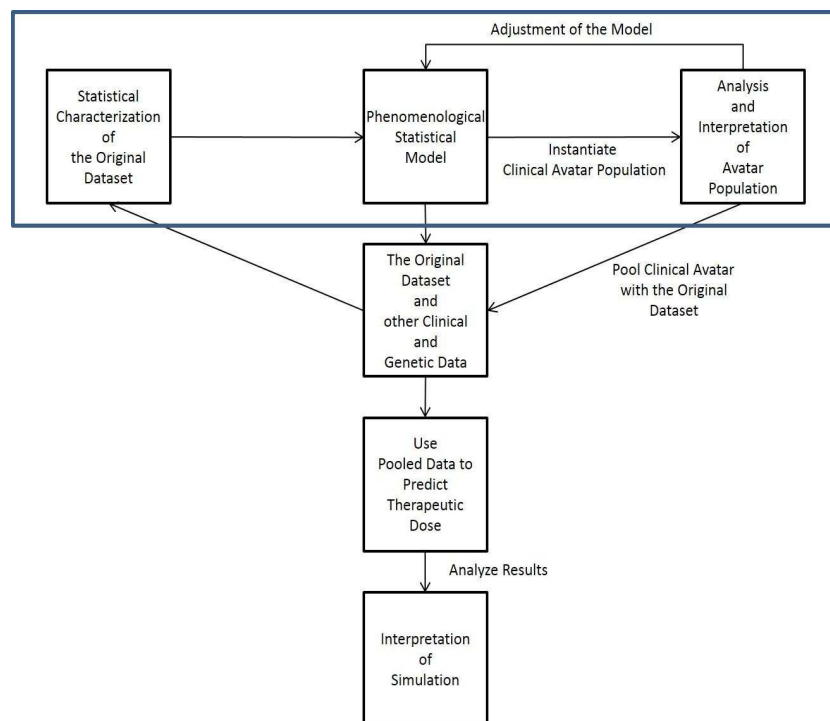
Figure 3.20. The Applied mathematical modeling schemata to generate clinical avatars consists of an iterative process of model development, testing, validation, refinement and repeat.

3.4.1.) Patient Populations

Enormous number of clinical trials have been conducted to test the value of genotyping on warfarin dosing and treatment. In fact, 364 clinical trials have been or are being conducted to test the accuracy of previous dosing algorithms, construct new dosing algorithms, or test the value of genetic variants in warfarin dosing (clinicaltrials.gov). In those costly clinical trials, tens of thousands of patients have been or are being recruited. The variety of trials, objectives, subject populations and results demonstrate a complex problem that has no obvious solution so far (~100 warfarin trials are still open as of June 1, 2014).

As mentioned, more than 40 warfarin prediction algorithms that use patient specific data to predict therapeutic warfarin dosing have been published in the last three decades. This collection of algorithms contain a number of

physiological, genetic, clinical lab, and clinical care variables used in one or

more algorithms. We have recorded each variable's nomenclature, data

constraints, range of values, and data type (Table 3.6). It is a source of

defining which variable (clinical, genetic, personal or family) should be

included in the algorithms.

Table 3.6. General characteristics of the variables used in most of the warfarin dosing algorithms.

Field	Units	Description	Format	Constraint	Example
Race	n/a	race	Controlled Dictionary	{White, African-American, Native-American, Asian, Pacific-Islander, Other, Unknown}	Other
Age	years	age in years	Integer	13 <= x <= 94	25
Gender	n/a	gender	Controlled Dictionary	{M, F}	M
Height	in	Height in inches	Integer	56 <= x < = 82	65
Weight	lbs	Weight in pounds	Integer	100 <= x < = 308	165
BSA	m^2	Body Surface Area	Real Number	1.3 <= x <= 2.8	1.6
CYP2C9	n/a	CYP2C9 genotype	Controlled Dictionary	{*1/*1,*1/*2,*1/*3, *2/*2,*2/*3,*3/*3}	*1/*1
VKORC1	n/a	VKORC1 genotype	Controlled Dictionary	{A/A, A/B, B/B}	A/A
CYP2C9*2	n/a	CYP2C9*2 genotype	Controlled Dictionary	{C/C, C/T, T/T}	C/T
CYP2C9*3	n/a	CYP2C9*3 genotype	Controlled Dictionary	{A/A, A/C, C/C}	A/C
VKORC1 (1173)	n/a	VKORC1(1173) genotype	Controlled Dictionary	{C/C, C/T, T/T}	C/T
VKORC1	n/a	VKORC1(-	Controlled	{G/G, G/A, A/A}	G/A

(-1639)		1639) genotype	Dictionary		
DVT	n/a	Warfarin usage indicated for DVT/PE or not	Controlled Dictionary	{Y, N}	Y
Smoker	n/a	Smokes or not	Controlled Dictionary	{Y, N}	Y
Target INR	n/a	Desired INR	Real Number	x = 2.5	2.5
Amiodarone	n/a	Amiodarone use or not	Controlled Dictionary	{Y, N}	Y

Based on our simulation study design, available population dataset that will be used for training and validation of Clinical Avatar BNM and the required variables in the simulation model, demographic, clinical, and genetic characteristics of the dataset are extracted and calculated. Prior to this step, the datasets are carefully vetted to determine the quality and quantitative properties.

3.4.2.) Parameters

For each study population, we extract and record the clinically and physiological valid ranges for each variable and create a representative statistical correlation of the variables important to the study. As explained in section 3.4.1, these characteristics are crucial to create clinical avatars model and simulation and PK/PD models. This information along with published statistics, correlations, and clinical associations will be used to define a clinical avatar statistical data model representative of a hypothetical population of warfarin patients. Then the clinical avatar simulation framework is used to generate clinical avatar populations. Each clinical avatar will reflect a hypothetical patient's medical record and the collection of avatars (the clinical

avatar population) will adhere to the prescribed statistics and inter-variable correlation structure prescribed by the clinical avatar model.

As an example, Table 3.7 presents the initial statistical distribution of the 5700 warfarin patient medical record datasets in PharmGKB (Whirl-Carrillo, 2012). The statistical results displayed in the table show the first order analysis (mean and standard deviation) of the variables assuming independence. The far right column provides a p-value for tests between the actual PharmGKB data and representative simulated clinical avatars.

Table 3.7. Statistical characteristics of the PharmGKB warfarin dataset versus its clinical avatar population.

Parameter	Actual PharmGKB Patients (n=5700)	Clinical Avatars (n = 20,000)	P-Value
Age			
<18	0.18% (10)	0.13% (26)	
18 – 24	1.3% (75)	1.2% (235)	
25 – 44	9.9% (559)	9.8% (1,957)	0.75
45 – 64	36% (2,040)	36.4% (7,282)	
65 – 94	52.5% (2,974)	52.5% (10,500)	
Gender by age			
<18	M: 30% (3)	M: 34.6% (9)	
	F: 70% (7)	F: 65.4% (17)	
18 – 24	M: 42.7% (32)	M: 47.2% (111)	
	F: 57.3% (43)	F: 52.7% (124)	
25 – 44	M: 49.9% (279)	M: 50.6% (990)	0.89
	F: 50.1% (280)	F: 49.4% (967)	
45 – 64	M: 60% (1225)	M: 59.4% (4,324)	
	F: 40% (815)	F: 40.6% (2,958)	
65 – 94	M: 59.3% (1,855)	M: 59.4% (6,353)	
	F: 40.7% (1,272)	F: 40.6% (4,344)	
Race			
White	54.8% (3,122)	54.2% (10,835)	
African American	8.1% (462)	7.9% (1,583)	
Native American	0% (0)	0%	
Asian	28.7% (1,634)	29.7% (5,936)	0.51
Pacific Islander	0% (0)	0%	
Other	0% (0)	0%	
Unknown	8.4% (482)	8.2% (1,646)	
Height (in)			
Mean	66.11 ± 4.3	66.50 ± 4	1.2e-8
Min	49	49	
Max	80	80	

Weight (lbs)			
Mean	171.58 ± 48.2	173.51 ± 27.87	7.7e-31
Min	66	92	
Max	524	290	
Smoker			
White	14.4% (324)	14.3% (1,552)	
African American	20.8% (91)	20.9% (332)	2.4e-11
Native American	0% (0)	0% (0)	
Asian/Pac Islander	6.4% (18)	5.7% (340)	
Other/Unknown	6.5% (16)	5.7% (94)	
Amiodarone			
Yes	4.5% (258)	4.6% (921)	
No	95.5% (5,442)	95.4% (19,079)	
DVT			
Yes	16.4% (817)	16% (3,203)	
No	83.6% (4,191)	84% (16,797)	
VKORC1			
A/A	52.2% (1,245)	52% (10,404)	
A/B	25.8% (614)	26.3% (5,261)	0.83
B/B	22.0% (525)	21.7% (4,335)	
CYP2C9			
*1/*1	74.9% (4,155)	75.4% (15, 079)	
*1/*2	13.4% (742)	13.4% (2,676)	
*1/*3	9% (501)	8.8% (1,756)	0.81
*2/*2	1% (58)	1% (194)	
*2/*3	1.3% (72)	1.1% (227)	
*3/*3	0.4% (22)	0.3% (68)	

As mentioned in the previous sections, we use PK/PD models to simulate anticoagulation medication response (e.g., INR). The PK/PD model predicts treatment outcomes depending on the medication, its metabolism, clearance and physical properties. As an example, 2 different clinical avatars (that represent 2 patients) may have different outcomes on the same medication-dosing algorithm and -protocol plan. In addition, the clinical avatars may have very different outcomes for 2 different medication-protocol plans or even for the same medication but different dosing protocols. As explained in section 3.2.7, we have implemented a PK/PD model for daily warfarin dosing whose PK model uses a 2-compartment model with first order input and first order elimination and its PD model affects using a 2-chain transit compartment

model. The predictive variables for this PK/PD model are: age, CYP2C9 and VKORC1 genotypes. So, these three characteristics of the clinical avatars are used to run the PK/PD model and predict INRs.

3.5.) Outcome Metrics

In anticoagulation therapy studies, a number of outcome metrics are used to measure the quality of therapy. They generally are categorized into two group; primary and secondary.

3.5.1.) Primary Outcome Metrics

The main primary outcome is the percentage of time in therapeutic range (TTR) of international normalized ratio (INR). TTR could be calculated using different methods explained in section 3.3.2. TTR is the most popular and widely used quality measure to monitor warfarin management. During the process of anticoagulation which is usually a long term one, most risk factors for bleeding or thromboembolic events such as age and underpinning comorbid conditions cannot be changed and the only modifiable factor that can be improved to avoid such complications is TTR (Levine, 1998).

3.5.2.) Secondary Outcome Metrics

This group of outcomes are categorized in the two following groups; principal secondary outcomes and non-principal secondary outcomes. The secondary outcomes are as follows:

- Occurrence of INR >4 or major clinical events in the first 4 weeks. This composite outcome measure is a principal secondary outcome measure and is defined as any INR of 4 or more, major bleeding, or thromboembolism in the first 4 weeks. The major clinical outcomes to be included in this measure

are major bleeding and thromboembolism events. The reason this measure is an important outcome measure is that many studies have shown that there is a significant correlation between INR >4 and increase risk of bleeding (van der Meer, 1993).

- Time to first therapeutic INR. This is measure defined as the first INR that is between the therapeutic ranges of INR depending on the indication of anticoagulation therapy. Even though it looks that this measure is an important one but there is no evidence showing a correlation between this measure and improved clinical outcomes of anticoagulation.

- Time to the determination of a maintenance dose. This is an outcome measure which is defined as the time to two consecutive INR measurements in therapeutic range without any change in dose, measured at least 1 week apart (Kimmel, 2013).

- Time to an adverse event. This measure is defined as the time to occurrence of adverse events ranging from minor bleedings and thromboembolism (TIA) to major bleedings and thromboembolism as well as death in specific time frames (e.g., 4 weeks and 3 months). Bleeding is the most serious complication of the use of oral anticoagulation in the prevention and treatment of thromboembolic complications. The definition of this event as used in major clinical trials is the following one adopted from the Italian Study on Complications of Oral Anticoagulant Therapy (Palareti, 1996): Major bleeding is classified as: "fatal (death due to hemorrhage); intracranial (documented by imaging), ocular (with blindness), articular, or retroperitoneal; if surgery or angiographic intervention was required to stop bleeding; and if bleeding led to hemoglobin reduction of 2 g/dL or more and/or need for transfusion of two or

more blood units. Minor bleeding is all cases of bleeding not classified as major. Non-relevant (small) bleeding is bruising, small ecchymosis or epistaxis, occasional hemorrhagic bleeding, or microscopic hematuria."

3.6.) References:

American Geriatrics Society Clinical Practice Guidelines (AGSCP Guidelines). (2000). The use of oral anticoagulants (warfarin) in older people. Journal of American Geriatrics Society, 48:224-227.

Anderson, J. L., Horne, B. D., Stevens, S. M., Grove, A. S., Barton, S., Nicholas, Z. P., et al. (2007). Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. Circulation, 116(22), 2563-2570.

Ansell, J., Hirsh, J., Hylek, E., Jacobson, A., Crowther, M., & Palareti, G. (2008). Pharmacology and management of the vitamin K antagonists: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). Chest, 133(6 Suppl), 160S-198S.

Ansell, J., Hirsh, J., Poller, L., Bussey, H., Jacobson, A., & Hylek, E. (2004). The pharmacology and management of the vitamin K antagonists: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest, 126(3 Suppl), 204S-233S.

Beal, S.L. & Sheiner, L.B. (1994). NONMEM Users Guide, NONMEM Project Group, University of California, San Francisco, CA.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics, 77(2), 81-97.

Berry, S. M., Carlin, B. P., Lee, J. J., & Muller, P. (2010). Bayesian adaptive methods for clinical trials. CRC Press.

Bielza, C., & Larrañaga, P. (2014). Discrete Bayesian network classifiers: a survey. ACM Computing Surveys (CSUR), 47(1), 5.

Broom, B. M., Do, K. A., & Subramanian, D. (2012). Model averaging strategies for structure learning in Bayesian networks with limited data. BMC bioinformatics, 13(Suppl 13), S10.

Burmester, J. K., Berg, R. L., Yale, S. H., Rottscheit, C. M., Glurich, I. E., Schmelzer, J. R., et al. (2011). A randomized controlled trial of genotype-based Coumadin initiation. Genetics in medicine : official journal of the American College of Medical Genetics, 13(6), 509-518.

Caraco, Y., Blotnick, S., & Muszkat, M. (2008). CYP2C9 genotype-guided warfarin prescribing enhances the efficacy and safety of anticoagulation: a prospective randomized controlled study. Clinical pharmacology and therapeutics, 83(3), 460-470.

Cavallari, L. H., & Perera, M. A. (2012). The future of warfarin pharmacogenetics in under-represented minority groups. Future cardiology, 8(4), 563-576.

CER Resources. (2014). Retrieved January 1, 2014 from https://ccts.osu.edu/education-and-training-programs/resource-room

Chickering, D. M. (2003). Optimal structure identification with greedy search. The Journal of Machine Learning Research, 3, 507-554.

Clyde, M. A., & Lee, H. K. (2001). Bagging and the Bayesian bootstrap. In Artificial Intelligence and Statistics (Vol. 169174).

Coumadin tablets (Warfarin sodium tablets, USP) Crystalline coumadin for injection (Warfarin sodium for injection, USP) [package insert]. (2011). Princeton, NJ: Bristol-Myers Squibb Company. Retrieved March 23, 2014 from Available from: http://dailymed.nlm.nih.gov/dailymed/lookup .cfm?setid=c0f238ff-2804-4150-9766-22a5111b4e74.

Daniels, P. R. (2005). Therapy Insight: management of urology patients taking long-term warfarin anticoagulation therapy. Nature clinical practice. Urology, 2(7), 343-350.

Dash, D., & Druzdzel, M. J. (2002, August). Robust independence testing for constraint-based learning of causal structure. In Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence (pp. 167-174). Morgan Kaufmann Publishers Inc..

Domingos, P. (1997). Why Does Bagging Work? A Bayesian Account and its Implications. In KDD, pp. 155-158.

Fiot, C., Saptawati, G. P., Laurent, A., & Teisseire, M. (2008, January). Learning Bayesian network structure from incomplete data without any assumption. In Database Systems for Advanced Applications (pp. 408-423). Springer Berlin Heidelberg.

Flores, M. J., Gámez, J. A., Martínez, A. M., & Puerta, J. M. (2011). Handling numeric attributes when comparing Bayesian network classifiers: does the discretization method matter?. Applied Intelligence, 34(3), 372-385.

Friedman, N. (1997, July). Learning belief networks in the presence of missing values and hidden variables. In ICML (Vol. 97, pp. 125-133).

Friedman, N., Goldszmidt, M., & Wyner, A. (1999, July). Data analysis with Bayesian networks: A bootstrap approach. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 196-205). Morgan Kaufmann Publishers Inc..

Fushiki, T. (2010). Bayesian bootstrap prediction. Journal of Statistical Planning and Inference, 140(1), 65-74.

Elting, L. S., Cooksley, C., Bekele, B. N., Frumovitz, M., Avritscher, E. B., Sun, C., et al. (2006). Generalizability of cancer clinical trial results: prognostic differences between participants and nonparticipants. Cancer, 106(11), 2452-2458.

Gage, B. F., Eby, C., Johnson, J. A., Deych, E., Rieder, M. J., Ridker, P. M., et al. (2008). Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. Clinical pharmacology and therapeutics, 84(3), 326-331.

Garcia, D., Regan, S., Crowther, M., Hughes, R. A., & Hylek, E. M. (2005). Warfarin maintenance dosing patterns in clinical practice: implications for safer anticoagulation in the elderly population. Chest, 127(6), 2049-2056.

Garcia, S., Luengo, J., Sáez, J. A., López, V., & Herrera, F. (2013). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. Knowledge and Data Engineering, IEEE Transactions on, 25(4), 734-750.

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. Neural Computing and Applications, 19, 263-282.

Geiger, D. and Heckerman, D. (1994). Learning gaussian networks. In Proceedings of the Tenth international conference on Uncertainty in artificial intelligence, 235-243. Morgan Kaufmann Publishers Inc..

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. CRC press.

Ghahramani, Z., & Jordan, M. I. (1994). Supervised learning from incomplete data via an EM approach. Advances in Neural Information Processing Systems 6.

Gibaldi, M., & Perrier, D. (1982). Pharmacokinetics Second Edition Revised and Expanded: Marcel Dekker, New York.

Glasgow, R. E., Nelson, C. C., Strycker, L. A., & King, D. K. (2006). Using RE-AIM metrics to evaluate diabetes self-management support interventions. American journal of preventive medicine, 30(1), 67-73.

Guyatt, G. H., Akl, E. A., Crowther, M., Gutterman, D. D., & Schuunemann, H. J. (2012). Executive summary: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest, 141(2 Suppl), 7S-47S.

Hamberg, A. K., Dahl, M. L., Barban, M., Scordo, M. G., Wadelius, M., Pengo, V., et al. (2007). A PK-PD model for predicting the impact of age, CYP2C9, and VKORC1 genotype on individualization of warfarin therapy. Clinical pharmacology and therapeutics, 81(4), 529-538.

Han, J., & Kamber, M. (2006). Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann.

Hutten, B. A., Prins, M. H., Redekop, W. K., Tijssen, J. G., Heisterkamp, S. H., & Buller, H. R. (1999). Comparison of three methods to assess therapeutic quality control of treatment with vitamin K antagonists. Thrombosis and haemostasis, 82(4), 1260-1263.

Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Statistics in medicine, 26(29), 5320-5334.

Johnson, J. A., Gong, L., Whirl-Carrillo, M., Gage, B. F., Scott, S. A., Stein, C. M., et al. (2011). Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. Clinical pharmacology and therapeutics, 90(4), 625-629.

Juni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ, 323(7303), 42-46.

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. The Journal of Machine Learning Research, 8, 613-636.

Kalisch, M., Mächler, M., Colombo, D., Hauser, A., Maathuis, M. H., & Bühlmann, P. (2014). More Causal Inference with Graphical Models in R Package pcalg.

Kessler, R., & Glasgow, R. E. (2011). A proposal to speed translation of healthcare research into practice: dramatic change is needed. American journal of preventive medicine, 40(6), 637-644.

Khoury, M. J., Gwinn, M., Yoon, P. W., Dowling, N., Moore, C. A., & Bradley, L. (2007). The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?. Genetics in medicine, 9(10), 665-674.

Kimmel, S. E., French, B., Kasner, S. E., Johnson, J. A., Anderson, J. L., Gage, B. F., et al. (2013). A pharmacogenetic versus a clinical algorithm for warfarin dosing. The New England journal of medicine, 369(24), 2283-2293.

Klein, T. E., Altman, R. B., Eriksson, N., Gage, B. F., Kimmel, S. E., Lee, M. T., et al. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. The New England journal of medicine, 360(8), 753-764.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI (Vol. 14, No. 2, pp. 1137-1145).

Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering, 32(1), 47-58.

Levine, M. N., Raskob, G., Landefeld, S., & Kearon, C. (1998). Hemorrhagic complications of anticoagulant treatment. [Review]. Chest, 114(5 Suppl), 511S-523S.
Li, T., Chang, C. Y., Jin, D. Y., Lin, P. J., Khvorova, A., & Stafford, D. W. (2004). Identification of the gene for vitamin K epoxide reductase. Nature, 427(6974), 541-544.

Lindh, J. D., Lundgren, S., Holm, L., Alfredsson, L., & Rane, A. (2005). Several-fold increase in risk of overanticoagulation by CYP2C9 mutations. Clinical pharmacology and therapeutics, 78(5), 540-550.

Loeliger, E. A. (1985). Laboratory control, optimal therapeutic ranges and therapeutic quality control in oral anticoagulation. Acta haematologica, 74(3), 125-131.

Lorbert, A., Blei, D. M., Schapire, R. E., & Ramadge, P. J. (2012). A Bayesian Boosting Model. arXiv preprint arXiv:1209.1996.

Margaritias D, and Thrun S. (2000). Bayesian network induction via local neighborhoods. In advances in Neural Information Processing Systems 12 (NIPS-1999). MIT Press, 505-511.

McWilliam AL, R.; Nardinelli, C. (2006). Health care savings from personalized medicine using genetic testing: the case of warfarin. Working Paper: AE1-Brookings Joint Center for Regulatory Studies.

Monteith, K., Carroll, J. L., Seppi, K., & Martinez, T. (2011, July). Turning Bayesian model averaging into Bayesian model combination. In Neural Networks (IJCNN), The 2011 International Joint Conference on (pp. 2657-2663). IEEE.

Monti, S., & Cooper, G. F. (1998, July). A multivariate discretization method for learning Bayesian networks from mixed data. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (pp. 404-413). Morgan Kaufmann Publishers Inc..

Palareti, G., Leali, N., Coccheri, S., Poggi, M., Manotti, C., D'Angelo, A., & Musolesi, S. (1996). Bleeding complications of oral anticoagulant treatment: an inception-cohort, prospective collaborative study (ISCOAT). The Lancet,348(9025), 423-428.

Pearl, J. (2000). Causality: models, reasoning and inference. Cambridge, UK: Cambridge University Press.

Pirmohamed, M. (2006). Warfarin: almost 60 years old and still causing problems. British journal of clinical pharmacology, 62(5), 509-511.

Ramoni, M., & Sebastiani, P. (2001). Robust learning with missing data. Machine Learning, 45(2), 147-170.

Rieder, M. J., Reiner, A. P., Gage, B. F., Nickerson, D. A., Eby, C. S., McLeod, H. L., et al. (2005). Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. The New England journal of medicine, 352(22), 2285-2293.

Rosendaal, F., Cannegieter, S., Van Der Meer, F., & Briet, E. (1993). A method to determine the optimal intensity of oral anticoagulant therapy. Thrombosis and haemostasis, 69(3), 236-239.

Rost, S., Fregin, A., Ivaskevicius, V., Conzelmann, E., Hortnagel, K., Pelz, H. J., et al. (2004). Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature, 427(6974), 537-541.

Sanderson, S., Emery, J., & Higgins, J. (2005). CYP2C9 gene variants, drug dose, and bleeding risk in warfarin-treated patients: a HuGEnet systematic review and meta-analysis. Genetics in medicine : official journal of the American College of Medical Genetics, 7(2), 97-104.

Schelleman, H., Chen, J., Chen, Z., Christie, J., Newcomb, C. W., Brensinger, C. M., et al. (2008). Dosing algorithms to predict warfarin maintenance dose in Caucasians and African Americans. Clinical pharmacology and therapeutics, 84(3), 332-339.

Schmitt, L., Speckman, J., & Ansell, J. (2003). Quality assessment of anticoagulation dose management: comparative evaluation of measures of time-in-therapeutic range. Journal of thrombosis and thrombolysis, 15(3), 213-216.

Schmidt M., Niculescu-Mizil A., Kevin Murphy K. (2007). Learning graphical model structure using l1-regularization paths. In Proceedings of the National Conference on Artificial Intelligence, volume 22, page 1278. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Annals of internal medicine, 152(11), 726-732.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search. 2nd edition. New York, NY: MIT Press.

Spirtes, P., & Meek, C. (1995). Learning Bayesian networks with discrete variables from data. In KDD, vol. 1, pp. 294-299. 1995.

Tsamardinos, I., Brown, L.E., & Aliferis, C.F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. Machine learning, 65(1):31, 78.

The Patient Protection and Affordable Care Act of 2010 (PPACA). (2010). 111–1482010. Retrieved August 10, 2014, from http://www.gpo.gov/fdsys/pkg/BILLS-111hr3590enr/pdf/BILLS-111hr3590enr.pdf

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. Machine learning, 65(1), 31-78.

Tunis, S. R., Stryer, D. B., & Clancy, C. M. (2003). Practical clinical trials: increasing the value of clinical research for decision making in clinical and

health policy. JAMA : the journal of the American Medical Association, 290(12), 1624-1632.

Upton, R. N., & Mould, D. R. (2014). Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development: Part 3-Introduction to Pharmacodynamic Modeling Methods. CPT: pharmacometrics & systems pharmacology, 3(1), e88.

van der Meer, F. J., Rosendaal, F. R., Vandenbroucke, J. P., & Briet, E. (1993). Bleeding complications in oral anticoagulant therapy. An analysis of risk factors. Archives of internal medicine, 153(13), 1557-1562.

Veenstra DL, You JHS, Rieder MJ, et al. (2005). Association of Vitamin K epoxide reductase complex 1 (VKORC1) variants with warfarin dose in a Hong Kong Chinese patient population. Pharmacogenet Genomics, 15:687-691.

Wadelius, M., Chen, L. Y., Downes, K., Ghori, J., Hunt, S., Eriksson, N., et al. (2005). Common VKORC1 and GGCX polymorphisms associated with warfarin dose. The pharmacogenomics journal, 5(4), 262-270.

Wadelius, M., Sorlin, K., Wallerman, O., Karlsson, J., Yue, Q. Y., Magnusson, P. K., et al. (2004). Warfarin sensitivity related to CYP2C9, CYP3A5, ABCB1 (MDR1) and other factors. The pharmacogenomics journal, 4(1), 40-48.

Wang, M., Lang, X., Cui, S., Fei, K., Zou, L., Cao, J., et al. (2012). Clinical application of pharmacogenetic-based warfarin-dosing algorithm in patients of Han nationality after rheumatic valve replacement: a randomized and controlled trial. International journal of medical sciences, 9(6), 472-479.

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., et al. (2012). Pharmacogenomics knowledge for personalized medicine. Clinical pharmacology and therapeutics, 92(4), 414-417.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. Evolutionary Computation, IEEE Transactions on, 1(1), 67-82.

Wysowski, D. K., Nourjah, P., & Swartz, L. (2007). Bleeding complications with warfarin use: a prevalent adverse effect resulting in regulatory action. Archives of internal medicine, 167(13), 1414-1419.

Yan, X., & Kimko, H. H. (2013). Pharmacokinetic and pharmacodynamic modeling of drug absorption and delivery.

Yan, X., & Kimko, H. H. (2013). Pharmacokinetic and pharmacodynamic modeling of drug absorption and delivery Therapeutic Delivery Methods: A Concise Overview of Emerging Areas (pp. 24-38): Future Science Ltd.

Yasar, U., Eliasson, E., Dahl, M. L., Johansson, I., Ingelman-Sundberg, M., & Sjoqvist, F. (1999). Validation of methods for CYP2C9 genotyping: frequencies of mutant alleles in a Swedish population. Biochemical and biophysical research communications, 254(3), 628-631.

Yuan, H. Y., Chen, J. J., Lee, M. T., Wung, J. C., Chen, Y. F., Charng, M. J., et al. (2005). A novel functional VKORC1 promoter polymorphism is associated with inter-individual and inter-ethnic differences in warfarin sensitivity. Human molecular genetics, 14(13), 1745-1751.

Appendix
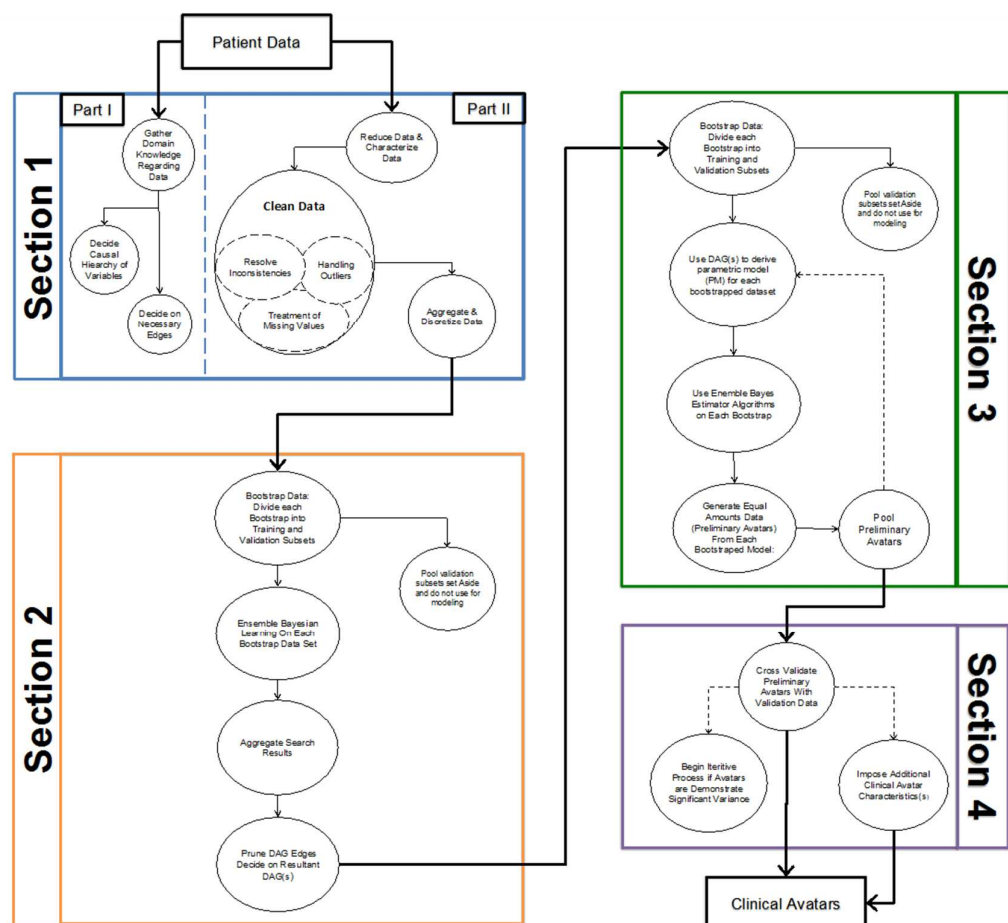
Figure 3.21. Our Complete Bayesian Network Modeling Pipeline for Clinical Avatars.

Chapter 4: A Comparative Effectiveness Anticoagulation/Anticlotting Research Database: WiAD

4.1.) Introduction

The Patient-Centered Outcomes Research Institute (PCORI) has proposed and approved five national priorities for research (Shelby, 2012). The goal of the fifth priority "Accelerating patient-centered outcomes research and methodology" is "Improving the nation's capacity to conduct patient-centered outcomes research, by building data infrastructure, improving analytic methods, and training researchers, patients and other stakeholders to participate in this research." The PCORI's "National Priorities for Research and Research Agenda" encompasses a number of prioritized research areas. Secondary use of large diverse healthcare EMR data for patient-centered comparative effectiveness research (PC-CER) is highlighted under the fifth priority as "The Research that determines the validity and efficiency of data sources commonly used in PCOR." The offered examples are: research that seeks to improve the volume, completeness, comprehensiveness, accuracy, and efficiency of use of clinical data collected across healthcare systems, clinical data networks, registries, or payer databases, and the utility of this data for conducting longitudinal studies of patient outcomes; research that explores the potential of large clinical data networks to support PCOR; or research that develops and promotes the utility, performance, and efficiency of large clinical data networks or registries for supporting patient-centered outcomes research for patients with rare diseases. This agenda encourages conducting a spectrum of PC-CER studies including health care disparity in part by seeking evidence of treatments' effectiveness across various populations.

American Medical Informatics Association (AMIA) has defined secondary use of data (Safran, 2007) as "non-direct care use of personal health information (PHI) including but not limited to analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities." Basically, secondary use of data implies use of health data for any purpose or activity other than direct healthcare delivery such as quality and safety measurement, clinical and translational research and improvement public health. Quality EMR data is central to any secondary use of the data and is essential to the process of decision-making and providing good quality patient care (Cruz-Correia, 2010). Secondary use of EMR as a means of retrospective analysis of health data has potentials to accelerate knowledge translation in healthcare and constitutes a significant part of clinical research. Currently, secondary use of clinical data is still at its early stage (Prokosch, 2009). National initiatives have been established to extend and facilitate secondary use of EMR to support clinical research (CDRNs, 2014).

Aurora Health Care (AHC) is the largest health care system in Wisconsin operating 15 hospitals throughout the state with more than 3600 licensed beds, 172 physician clinic facilities, and several other health care related entities. It serves about 1.2 million unique patients each year through 7.8 million patient encounters per year. AHC's EMR is the most comprehensive (by size, type of health care and period of time) digitized health care resource of Southeast Wisconsin's population capturing urban, suburban, and rural constituents of all racial, ethnic, and socioeconomic backgrounds. This

resource provides a unique opportunity to capture data and information vital to conducting retrospective and predictive PC-CER studies.

UWM and AHC have established a collaboration to conduct a series of research studies whose goal is to, using AHC electronic medical records (EMR), demonstrate that *in silico* pharmacogenetic PC-CERs based on actual EMR patient data provide predictive evidence useful in detecting "optimal" anticoagulation dosing protocols that reduce adverse drug responses, improve overall patient outcomes and reduce health disparities in the Aurora Milwaukee County patient population. One of the main steps in this collaboration is to build a longitudinal EMR-based anticoagulation research database. Such a database is an essential tool in conducting pharmacogenetic PC-CERs.

However, there are challenges in the re-use of the data captured and stored in EMR systems for comparative-effectiveness studies as they have not been originally been developed for research purposes. Some of the most popular barriers and challenges to the secondary use of EMR data are as follows: missing data, erroneous data, uninterpretable data, inconsistencies among providers and over time, and data stored in noncoded text notes (free text) (Bayley, 2013; Elkin, 2010). Considering the fact that the designed pharmacogenetic PC-CER studies by UWM and AHC required quality data, a rigorous multilayer and iterative process of extraction, transformation and loading (ETL) was designed and performed to ensure the resulted research database fits the use. In the following sections the ETL process and the research database are explained.

4.2.) Methods

The process of developing the anticoagulation research database consisted of an iterative process of extraction, transformation and loading by the two teams of Aurora Health Care (AHC) and University of Wisconsin-Milwaukee (UWM). The first step was to determine the scope of the desired data.

4.2.1.) Determination of Inclusion Criteria and Scope of Data

After a rigorous literature review, the UWM team developed a data model of the required patient's attributes and value sets for extracting the required data from AHC EMR data repository. Based on some practical facts, the data model was refined collaboratively by the UWM and AHC teams to be compatible with the data models and dictionaries of the heterogeneous AHC's databases (Table 4.1.). For example, even though the initial data model included the full lists of National Drug Code (NDC) codes for each of the included medications or associated ICD-10 codes were considered for clinical characteristics (e.g., medical indication and comorbidities), the final model were refined to include medication names and ICD-9 codes given that AHC data warehouse were not using NDC and ICD-10 codes.

Table 4.1. Patient's characteristics used to identify patients with evidence of anticoagulation/anticlotting treatment from AHC's EMR data warehouse.

Patient Characteristics	ICD-9 or CPT Codes
Demographics	
Gender	Male, Female, Unknown
Age	Year (Age >=18)
Patient's Zipcode	5-digit zipcode
Patient's County	
Patient's City	
Provider's Zipcode	5-digit zipcode
Smoking Status	

Race	
Height	Inch
Weight	Pound
Date of Visit	
Date of Death	
Medications	
Date Received/Date Prescribed	

Medication Type:
  - Coumadin (Warfarin)
  - Heparin
  - Ticlopidine (Ticlid)
  - Clopidogrel (Plavix)
  - Dipyridamole (Persantine)
  - Abciximab (ReoPro)
  - Eptifibatide (Integrilin)
  - Tirofiban (Aggrastat)
  - Dabigatran (Pradaxa)
	Name of the medications

Dosage	
Frequency	
Interacting Medications	
Date Received/Date Prescribed	

Medication Type:
  - Amiodarone (Cordarone, Pacerone)
  - Simvastatin (Zocor)
  - Fluvastatin (Lescol)
  - Lovastatin (Altocor, Altoprev, Mevacor)
  - Atrovastatin (Lipitor)
  - Rosuvastatin (Crestor)
  - Pravastatin (Pravachol)
  - Aspirin
	Name of the medications

Dosage	
Frequency	
Medical Indications	
Date of Indication	
Indication Type:	
- Orthopedic Procedures	CPT codes: 77.65, 77.66, 77.67, 77.69, 77.75, 77.76, 77.77, 77.79, 77.85, 77.86, 77.87, 77.95, 77.96, 77.97, 78.05, 78.06, 78.07, 78.09, 78.15, 78.16, 78.17, 78.19, 78.25, 78.27, 78.29, 78.35, 78.37, 78.39, 78.45, 78.46, 78.47, 78.49, 78.55, 78.56, 78.57, 78.59, 79.05, 79.06, 79.09, 79.15, 79.16, 79.19, 79.25, 79.26, 79.29, 79.35, 79.36, 79.39, 79.45, 79.46, 79.49, 79.55, 79.56, 79.59, 79.65, 79.66, 79.69, 79.75, 79.76, 79.79, 79.85, 79.86, 79.89, 80.25, 80.26, 80.27, 80.6, 80.75, 80.76, 81.00, 81.09, 81.30, 81.39, 81.40, 81.42, 81.43, 81.44, 81.45, 81.46, 81.47, 81.49, 81.51, 81.52, 81.53, 81.54, 81.55, 81.56, 81.57,

		81.59, 81.61, 81.62, 81.63, 81.64, 81.65, 81.66, 84.10, 84.11, 84.12, 84.13, 84.14, 84.15, 84.16, 84.17, 84.18, 84.19, 84.26, 84.27, 84.28, 84.45, 84.46, 84.47, 84.48, 84.51, 84.58, 84.59, 84.60, 84.61, 84.62, 84.63, 84.64, 84.65, 84.66, 84.67, 84.68, 84.69, 84.80, 84.81, 84.82, 84.83, 84.84, 84.85, 84.91
	- Deep vein thrombosis (DVT)	ICD-9 codes: 451.x, 453.x
	- Pulmonary embolism (PE)	ICD-9 codes: 415.1x
	- Atrial fibrillation	ICD-9 codes: 427.31
	- Atrial flutter	ICD-9 codes: 427.32
	- Atrial fibrillation and flutter	ICD-9 codes: 427.3
	- Stroke	ICD-9 codes: 433.x, 434.x, 435.x, 436, 437.1x, 437.9x, 438.x
	- Heart valve replacement	ICD-9 codes: V43.3
Lab Test Results		
	INR Test Result	Numeric
	Date of INR Test	
Adverse Events/Comorbidities		
	Date of Event/Comorbidity	
	Event Type:	
	- Deep vein thrombosis (DVT)	ICD-9 codes: 451.x, 453.x
	- Pulmonary embolism (PE)	ICD-9 codes: 415.1x
	- Stroke	ICD-9 codes: 433.x, 434.x, 435.x, 436, 437.1x, 437.9x, 438.x
	- Myocardial infarction	ICD-9 codes: 410.x
	- Bleeding	ICD-9 codes: 431.x, 432.x, 459.0, 578, 578.9, 784.7, 784.8, 786.30
	- Diabetes Mellitus	ICD-9 codes: 250.x
	- Heart Failure	ICD-9 codes: 428.x
	- Hypertension	ICD-9 codes: 401.x
	- Peripheral artery diseases	ICD-9 codes: 250.6, 443.x, 785.4
	- Atherosclerosis of aorta	ICD-9 codes: 440-0
	- Coronary artery disease	ICD-9 codes: 414.0x

4.2.1.1.) Extract, Transform, and Load (ETL) Process

The process of extraction, transformation and loading (ETL) of data is usually a multilayer, iterative one addressing both data and data models. In this project, the ETL process was done by both AHC and UWM LPHIG teams.

4.2.1.1.1.) AHC ETL Process

Once the Internal Review Board (IRB) approval was granted by both AHC and UWM and a Data Use Agreement was signed by two institutes, the AHC team, before delivering the data to UWM team, performed an internally developed ETL process to identify, extract, transform and load patient's records into a database. Using the data model depicted in Table 4.1, the AHC team constructed the extraction algorithms and search strategies to mine AHC's hospitals' and clinics' EMR data warehouses and retrieved the AHC EMR for patients treated with anticoagulation/anticlotting agents for the period of 2002 to 2012 and subsequently all post-treatment events from each patient. In the process of the transformation, the patient data was de-identified per IRB approval (allowing zipcode) by an AHC honest broker. This process resulted in the longitudinal data records of 157,450 patients including: gender, race, height, weight, age, day of visit, patient's zipcode, patient's city, provider's zipcode, smoking status, INR, medications received (day, dose, frequency), interacting medications (Amiodarone, Simvastatin, Fluvastatin, Lovastatin, Atrovastatin, Rosuvastatin, Pravastatin, Aspirin), medication indications (by ICD-9 codes for Deep Vein Thrombosis (DVT), Pulmonary embolism, Atrial fibrillation, Atrial flutter, Atrial fibrillation and flutter, Stroke, Heart valve replacement and CPT codes for Orthopedic surgery-hip or knee) and comorbidities (by ICD-9 codes for DVT, Pulmonary embolism, Stroke, Myocardial infarction, Bleeding). The AHC team loaded the data into an MS Access database ("Original Access Database"). The data model of the Access database is depicted in Figure 4.1.
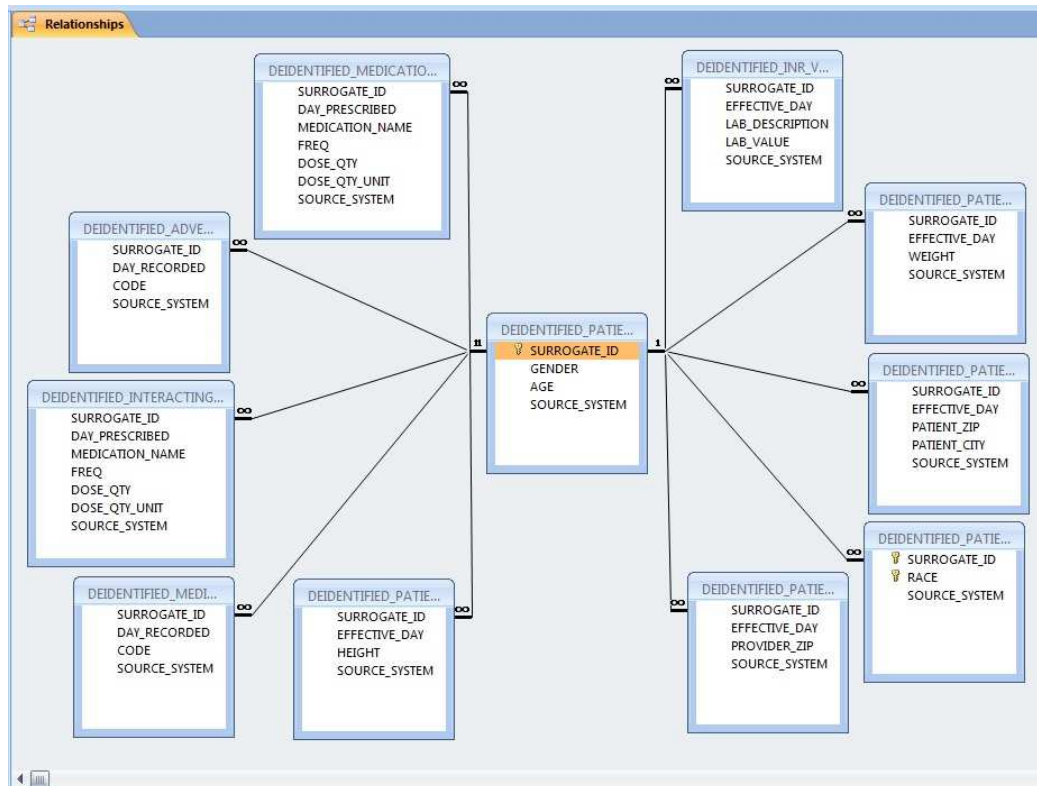
Figure 4.1. The Data Model of the Original Access Database Provided by AHC Team.

4.2.1.1.2.) LPHIG ETL Process

Once receiving the Original Access Database, the UWM LPHIG (Laboratory for Public Health Informatics and Genomics) team performed another iterative round of ETL process aiming to create the anticoagulation/anticlotting research database namely Wisconsin Anticoagulation Database (WiAD). The first round of the ETL process at this stage was to remove complete duplicate records from the Original Access Database's tables, identify primary and foreign keys across the tables and create a relational database in a MySQL Server which would be a "Base Database" for further processes. The Entity Relationship (ER) data model of the MySQL Base Database is depicted in Figure 4.2. This process was followed by another round of transformation including multiple steps of cleaning up the data and data validation and quality

control. In this process, few data dictionaries were developed for some of the attributes' value sets.

Data cleaning and quality control is an important task in the process of developing research databases especially when the extracted data come from heterogeneous data sources. Data cleaning is basically a process in which errors and inconsistencies in data are detected and removed.
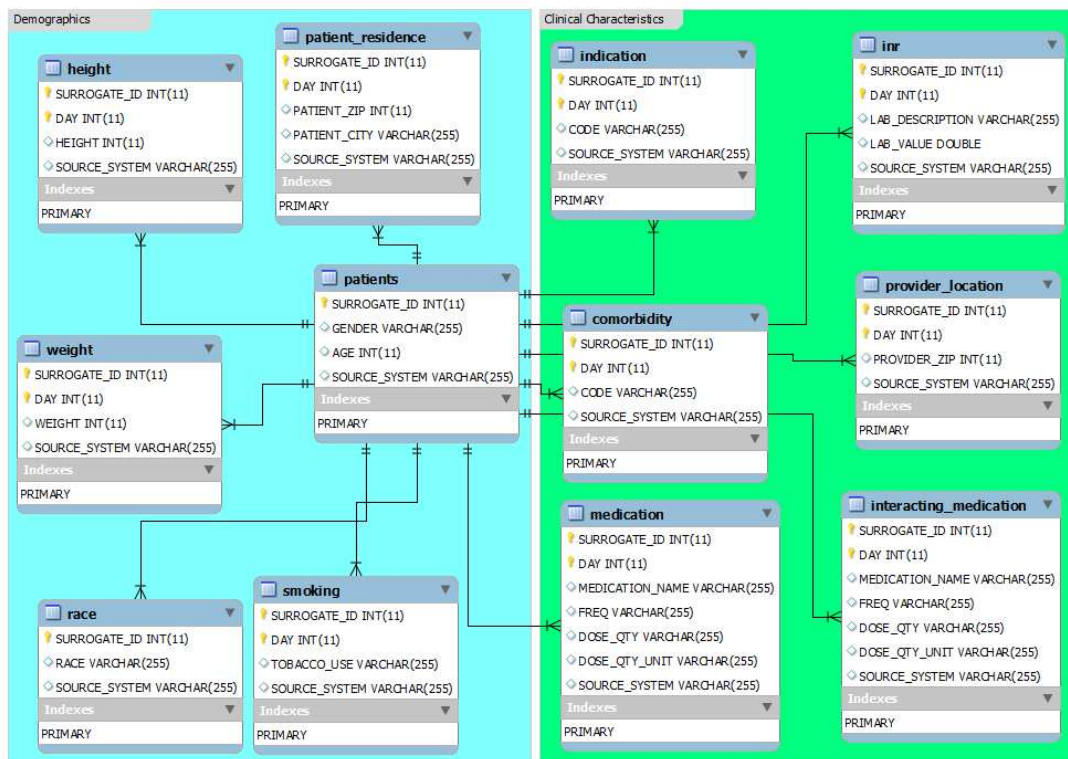
Figure 4.2. Entity Relationship (ER) Data Model of the Base MySQL Database.

Subsequent data cleaning, quality control and quality assurance included an iterative process of data parsing to detect irregularities; statistical analysis designed to test population-wide distributions and possible biases; refinement of inclusion and extraction data mining codes to address irregularities, possible missing data and detected biases; and ultimately, data transformation to produce a cohesive set of records capturing all available

medical records in a consistent format following Weiskopf (2013). According

to Weiskopf different dimensions of data quality are as follows:

- Completeness: "Is a truth about a patient present in the EMR?"

- Correctness: "Is an element that is present in the EMR true?"

- Concordance: "Is there agreement between elements in the EMR, or

between the EMR and another data source?"

Table 4.2. AHC dataset's data quality issues.

Scope/Problem		Original Data
Attribute	Missing values	Race = "UNKNOWN"
	Misspellings	Medication Name= "warfin"
	Awkward Abbreviations	Medication Frequency= "zUBC"
	Free Text Embedded values	Medication Name= "warfarin 2.5 mg 5 days a week and 2 mg two days a week"
	Miscoded values	Patient zip code= "WI"
	Incorrect values	Weight= -165
Record	Violated attribute dependency	City= "Milwaukee", zip code=99999
	Duplicated records (partial and complete)	ID=165, Day= 199; Medication= Warfarin 4 mg, Frequency= QOD; ID=165, Day= 199; Medication= Warfarin 4 mg, Frequency= daily
	Contradicting records	ID= 78, Day= 1101, Medication= Coumadin, Dose= 3 mg; ID= 78, Day= 1101, Medication= Coumadin, Dose= 4 mg

 - Plausibility: "Does an element in the EMR makes sense in light of other

knowledge about what that element is measuring?"

- Currency: "Is an element in the EMR a relevant representation of the patient state at a given point in time?"

We identified a number of data quality issues across the AHC dataset stored in the Base Database. Table 4.2 demonstrates the data quality issues of the AHC data at attribute and cross-attribute (record) levels.

To address these data quality issues across the AHC's structured and unstructured data, the team performed a number of data quality and cleaning tasks as detailed below.

- Age: Although the AHC data is longitudinal, there was only one Gender and one Age record for each unique subject stored in table "patients". The rate of missing data for age was very low (i.e., 1 subject). However, about 1% of the whole population (i.e., 2662 subjects) have age value of 0. Subjects with age of zero or missing age were excluded. Since our age inclusion criteria was subjects with the age of 18 years old or above, any subject with an age lower than 18 years old (i.e., 357 subjects) was excluded. The following table (Table 4.3.) shows the age distribution of subjects in the AHC Base Database. As depicted, the most populated age groups are 65-74 and 75-84 years old

Table 4.3. Age distribution of the subjects in the AHC Base Database.

Age Group (Years)	Percentage
18-24	1.09
25-34	2.46
35-44	5.81
45-54	14.11
55-59	9.92
60-64	10.62
65-74	22.22
75-84	22.70
>85	11.08

which is consistent with the fact that the prevalence of anticoagulation therapy increases as people get older and experience more chronic diseases.

- Gender: The gender values were either female or male. Less than 0.01% of the subjects had no gender information. The subjects with missing gender were excluded from the study population. Table 4.4 depicts the gender distribution of the subjects in the AHC Base Database compared to that of in Milwaukee county (MKE) and Wisconsin (WI). As a quality control measure, the age distribution of the AHC Base Database was compared with that of MKE and WI which did not show significant difference ($p > 0.05$).

Table 4.4. Gender Distributions across the three populations of AHC Base Database, Milwaukee County (MKE) and Wisconsin (WI).

Gender	AHC Base Database (%)	MKE (%)	WI (%)
Male	50.34	48.30	49.60
Female	49.65	51.70	50.40
Missing	0.01	NA	NA

- Race: In the AHC Base Database, each subject had one race record. Race is a key piece of information for our PC-CER studies. Only 52.85% of AHC Base Database's subjects had race information. About 0.01% of the subjects in table "race" (i.e., 1662 subjects) had no race information. The race values of the subjects were: White, Black or African American, Asian, American Indian or Alaskan Native, Native Hawaiian/Other Pacific Islander and Unknown. The following table (Table 4.5) depicts the racial distribution of the AHC Base Database's subjects who had identified race information versus the racial distributions across Milwaukee County (MKE) and Wisconsin (WI). As a quality control measure, the racial distribution of the AHC Base Database was compared with that of MKE and WI. The racial distributions of the Base Database was not significantly different from that of WI ($p > 0.05$).

Table 4.5. Race distribution across the AHC Base Database's subjects and
Milwaukee County's and Wisconsin's populations.

Race	AHC Base Database (%)	MKE (%)	WI (%)
White	90.41	65.9	88.4
Black or African American	8.52	27.0	6.5
American Indian and Alaska Native	0.27	0.9	1.1
Asian	0.79	3.6	2.4
Native Hawaiian or Other Pacific Islander	0.02	0.1	0.05
Subjects reporting two or more races	NA	2.5	1.6

- Height: In the AHC Base Database, each subject has longitudinal height

records with multiple height records for each measurement day which were

not on a regular basis. Each height record contained the attributes

{SURROGATE_ID, DAY, HEIGHT, SOURCE_SYSTEM}. Since height

measurement can vary within a specific time frame due to factors such as lack

of standard protocols, inaccurate measuring, and imprecise equipment set-up,
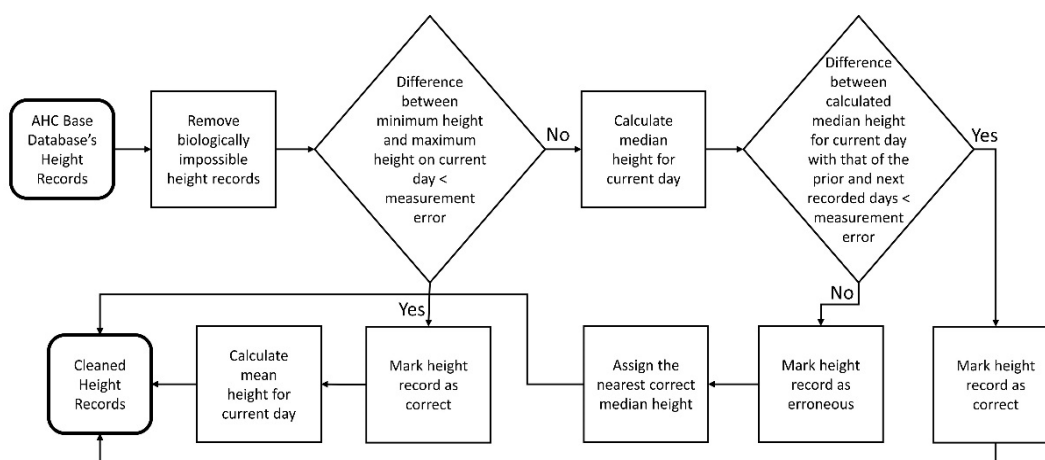
Figure 4.3. Height records cleaning algorithm.

we created a cleaning algorithm. Before applying the algorithm, first, each

height record was examined to identify and exclude any height record whose

HEIGHT attribute's value was missing. Then the height cleaning algorithm was applied (Figure 4.3).

According to the algorithm, we first removed all height values we considered biologically implausible (Census Bureau, 2012; Muthalagu, 2014). Accordingly, we set the plausible thresholds greater than 39.5 inches and less than 98.5 inches. Then we took the next steps of the algorithm. If all height records for a single subject in a single day differed (i.e., difference between minimum height and maximum height for the same day) by less than measurement error of 1.5 inches (Muthalagu, 2014), then all height records for that subject for that day were marked correct by the algorithm and the mean of the height records was calculated representing height value for the day. For all other days, median height for each day was calculated. The median height for each day was compared with that of the prior and next median; when the difference in medians was greater than 1.5 inches for both, the median height for that day was considered as potentially erroneous (Muthalagu, 2014). For the days that had an erroneous median height, the algorithm assigned the nearest correct median height.

The rate of height record completeness for the AHC Base Database was 43.17%. The subjects with height records had 3.07 height records on average. The average height of the population was 66.09 inches (±7.23).

- Weight: In the AHC Base Database, each subject has longitudinal weight records with multiple weight records for each measurement day which were not on a regular basis. Each weight record contained the attributes {SURROGATE_ID, DAY, WEIGHT, SOURCE_SYSTEM}. Since weight measurement can vary within a specific time frame due to factors such as

lack of standard protocols, inaccurate measuring, and imprecise equipment

set-up, we created a cleaning algorithm. Before applying the algorithm, first,

each weight record was examined to identify and exclude any weight record

whose WEIGHT attribute's value was missing. Then the weight cleaning

algorithm was applied (Figure 4.4).
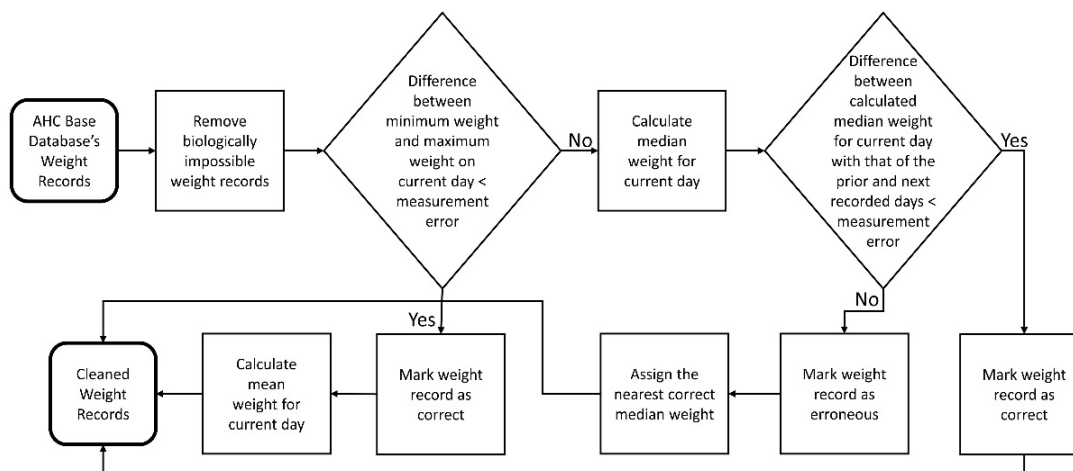
Figure 4.4. Weight records cleaning algorithm.

According to the algorithm, we first removed all weight values we considered

biologically implausible (Census Bureau, 2012). We set the plausible

thresholds greater than 70 pounds and less than 500 pounds. Then we took

the next steps of the algorithm. If all weight records for a single subject in a

single day differ (i.e., difference between minimum weight and maximum

weight for the same day) by less than measurement error of 10% (Maskin,

2010), then all weight records for that subject for that day are marked correct

by the algorithm and the mean of the weight records is calculated

representing weight value for the day. For all other days, median weight for

each day is calculated. The median weight for each day is compared with that

of the prior and next median; when the difference in medians is greater than

measurement error for both, the median weight for that day is considered as potentially erroneous. For the days that have an erroneous median weight, the algorithm assigns the nearest correct median weight.

The rate of weight record completeness for the AHC Base Database is 49.58%. The subjects with weight records had 7.39 weight records on average. The average weight of the population was 195.12 pounds (±50.1).

- Medications: In the AHC Base Database's table "medication", each subject has multiple medication records. Each medication record contained the attributes {SURROGATE ID, DAY, MEDICATION_NAME, FREQ, DOSE_QTY, DOSE_QTY_UNIT, SOURCE_SYSTEM}. In the table, the entire columns of {DOSE_QTY, DOSE_QTY_UNIT} were completely blank. The column {FREQ} was partially complete. The cells under column {MEDICATION_NAME} were populated with free-text. Before conducting any cleaning or quality control measures, we had to extract information from the free-text in cells under column {MEDICATION_NAME} and translate them into a structured format. Through a text analysis step, the required data from the free text were extracted and presented in normalized and consistent structured formats. We used a multistep process which involved parsing the free texts into their components, normalizing the identified components, and extraction of the required data elements (i.e., name of medication, dosage, unit, and route of administration). The rate of incompleteness of the resulted records was high. Table 4.6 depicts the distribution of medication records by each medication in the cleaned and controlled medication table and also presents an example of the rate of completeness (i.e., percentage of each medication's records with dosage information). 99.02% of the medications

records included one of the three medications of warfarin (37.85%), heparin
(36.24%) and clopidogrel (24.93%). Warfarin records which have dosage
information have the highest percentage of whole the records (11%). Although
the number of heparin records was higher than the number of clopidogrel
records, the percentage of clopidogrel records with dosage information out of
whole the records was higher than that of heparin.

Table 4.6. Distribution of number of subjects under each medication and the
rate of medication records' dosage information completeness.

Medication	Number of Subjects under Medication	Rate of the Records' Dosage Information Completeness
Warfarin	74,102	70.03%
Heparin	71,537	33.4%
Clopidogrel	61,517	68.03%
Dabigatran	1,793	89.18%
Dipyridamole	2,292	22.9%
Eptifibatide	2886	71.58%
Ticlopidine	434	73.5%
Abciximab	1,310	90.15%
Tirofiban	48	0%

The above table indicates that a big number of subjects have received more
than one medication. The following table (Table 4.7) shows the number of
subjects who have been under treatment with 1 to 6 medications.

Table 4.7. Distribution of subjects by number of medications.

No. of Medications	1	2	3	4	5	6
No. of Subjects	75282	16441	3049	358	20	1

- Patient Residence: In the AHC Base Database's table "patient_residence", each subject has multiple residence records. Each residence record contained the attributes {SURROGATE_ID, DAY, PATIENT_ZIP, PATIENT_CITY, SOURCE_SYSTEM}. As a quality control measure, the recorded 5-digit zipcodes were examined against a standard zipcode dictionary to find any inconsistency between the recorded zipcodes and their recorded associated cities. It also helped find and exclude invalid numeric or non-numeric values which were reported as 5-digit zipcodes (e.g., 0, 99999, WI, MI, *R2Y1, *CV11, *). 83.73% of the whole AHC Base Database's population have at least one residence record. The partially (i.e., the same zipcode on the same day at two different "SOURCE_SYSTEM"s) and complete duplicated records were also excluded (46.36% of the records). After the above QC and cleaning process, the average number of residence records for each subject with residence records was 1.76 (range 1-4). Out of the subjects with residence record, 3.33% had more than one reported zipcode. The reported zipcodes were distributed across 44 states with the highest frequency in Wisconsin, Illinois, Michigan, and Florida. The subjects in Wisconsin were also distributed across the 72 counties of the state with the highest frequency in counties Milwaukee, Sheboygan, Racine, Waukesha, Walworth, and Kenosha. The distribution of the subjects across the states and the counties have been visualized on two separate interactive Google Maps at the following URLs:

https://www.google.com/fusiontables/DataSource?docid=1MPew8EhUQvoES
HZJP9RaykxHDm8ir0W7LUhYC4U#map:id=3

https://www.google.com/fusiontables/DataSource?docid=1OP6Wrdt56g6ZAt0j
Yuyt3i4nc6RbUBFElqVtvnU

The following figure demonstrates a snapshot of the map displaying the

subject distribution across the counties of Wisconsin.
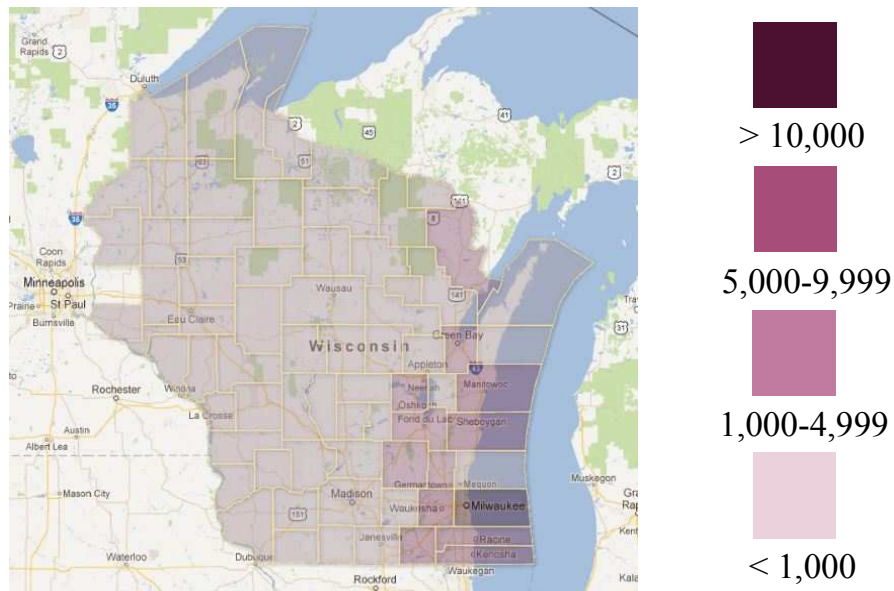
Figure 4.5. Distribution of AHC Base Database's subjects across
Wisconsin counties.

- Provider Location: In the AHC base database's table "provider_location",

each subject has multiple provider location's records. Each provider location's

record contained the attributes {SURROGATE ID, DAY, PROVIDER_ZIP,

SOURCE_SYSTEM}. As a quality measure, the recorded 5-digit zipcodes

were examined against a standard zipcode dictionary to find and exclude any

invalid recorded zipcodes. 54.10% of the whole AHC Base Database's

population have at least one provider location record. The partially (i.e., the

same zipcode on the same day at two different "SOURCE_SYSTEM"s) and

complete duplicated records were also excluded (3.93% of the records). The

average number of provider location records for subjects with provider

location records was 1.06 (range 1-4). Out of the subjects with provider

location record, 6.65% had more than one reported zipcode. The reported provider locations' zipcodes were distributed across 20 states with the highest frequency in Wisconsin and Illinois (97.75% and 1.15%, respectively). The provider locations in Wisconsin were also distributed across the 29 counties of the state with the highest frequency in counties Milwaukee and Brown (59.56% and 8.65%, respectively).

- Smoking: In the AHC Base Database's table "smoking", each subject has longitudinal smoking records with multiple smoking records for each record day which were not on a regular basis. Each smoking record contained the attributes {SURROGATE_ID, DAY, TOBACCO_USE, SOURCE_SYSTEM}. The TOBACCO_USE value fields were populated with a large number of different unstructured data values. To address this issue, a dictionary was created to refine and translate the recorded values to a set of well-defined ones (Table 4.8). Almost 68% of the Base Database's subjects had smoking status records.

Table 4.8. Smoking Status value dictionary.

Tobacco Use	Smoking Status	Definition
Current Smoker	YES	Currently Smoker
Former Smoker < or = to 12 months ago	NO	Not currently smoker
Former Smoker >12 months ago	NO	Not currently smoker
Never Smoker	NO	Never smoker
Unknown if Ever Smoked	NO	Unknown

- Indication: In the AHC Base Database's table "indication", each subject has multiple indication records. Each indication record contained the attributes {SURROGATE ID, DAY, CODE, SOURCE_SYSTEM}. As a quality measure, first, the reported codes were examined against the published ICD-9 CM 2011

codes (ICD-9-CM, 2014) to test their validity (i.e., in acceptable range). All of

the reported codes were in acceptable range (Table 4.9).

Table 4.9. ICD-9 Codes for indications of the AHC Base Database's subjects.

ICD-9-CM 2011 Diagnostic and Procedure Code Class	Indication codes reported at the AHC Base Database

Myocardial Infarction
  410 Acute myocardial infarction
  410.0 Of anterolateral wall
  410.1 Of other anterior wall
  410.2 Of inferolateral wall
  410.3 Of inferoposterior wall
  410.4 Of other inferior wall
  410.5 Of other lateral wall
  410.6 True posterior wall infarction
  410.7 Subendocardial infarction
  410.8 Of other specified sites
  410.9 Unspecified site
	410.01 410.02 410.11 410.12 410.21 410.22 410.32 410.41 410.51 410.72 410.81 410.82 410.91 410.92

Pulmonary Embolism
  415 Acute pulmonary heart disease
  415.1 Pulmonary embolism and infarction
	415.11 415.12 415.13 415.19

Cardiac Dysrhythmias
  427 Cardiac dysrhythmias
  427.3 Atrial fibrillation and flutter
	427.31 427.32

Cerebrovascular Disease
  432 Other and unspecified intracranial hemorrhage
    432.1 Subdural hemorrhage
  433 Occlusion and stenosis of precerebral arteries
    433.1 Occlusion and stenosis of carotid artery
    433.2 Occlusion and stenosis of vertebral artery
    433.8 Occlusion and stenosis of other specified precerebral artery
    433.9 Occlusion and stenosis of unspecified precerebral artery
  434 Occlusion of cerebral arteries
    434.1 Cerebral embolism
    434.9 Cerebral artery occlusion unspecified
  435 Transient cerebral ischemia
    435.1 Vertebral artery syndrome convert
    435.2 Subclavian steal syndrome convert
    435.3 Vertebrobasilar artery syndrome convert
    435.8 Other specified transient cerebral
	432.1 433.01 433.11 433.21 433.31 433.81 433.91 434.01 434.11 434.91 435.1 435.2 435.3 435.8 435.9 436 437.1 437.9 438.11 438.12 438.13 438.14 438.19 438.21 438.22 438.31 438.32 438.41 438.42 438.51 438.52 438.53 438.6 438.7 438.81 438.82 438.83 438.84 438.85 438.89 438.9

ischemias convert
   435.9 Unspecified transient cerebral
   ischemia convert
436 Acute, but ill-defined, cerebrovascular
disease
437 Other and ill-defined cerebrovascular
disease
   437.1 Other generalized ischemic
   cerebrovascular disease convert
   437.9 Unspecified cerebrovascular disease
   convert
438 Late effects of cerebrovascular disease
   438.0 Late effects of cerebrovascular
   disease, cognitive deficits
   438.1 Speech and language deficits
   438.2 Hemiplegia/hemiparesis
   438.3 Monoplegia of upper limb
   438.4 Monoplegia of lower limb
   438.5 Other paralytic syndrome
   438.6 Late effects of cerebrovascular
   disease, alterations of sensations
   438.7 Late effects of cerebrovascular
   disease, disturbances of vision
   438.8 Other late effects of cerebrovascular
   disease
   438.9 Unspecified late effects of
   cerebrovascular disease	

Gastrointestinal Hemorrhage
578 Gastrointestinal hemorrhage
   578.1 Blood in stool
   578.9 Hemorrhage of gastrointestinal tract,
   unspecified
	578.1 578.9

Hemorrhage in Head and Neck
784 Symptoms involving head and neck
   784.7 Epistaxis
   784.8 Hemorrhage from throat
	784.7 784.8

Hemorrhage in Respiratory System 786 Symptoms involving respiratory system and other chest symptoms 786.3 Hemoptysis	786.3

Phlebitis and Thrombophlebitis
451 Phlebitis and thrombophlebitis
   451.1 Phlebitis and thrombophlebitis of deep
   veins of lower extremities
   451.2 Phlebitis and thrombophlebitis of
   lower extremities, unspecified
   451.8 Phlebitis and thrombophlebitis of other
   sites
   451.9 Phlebitis and thrombophlebitis of
	451.11 451.19 451.2 451.81
	451.82 451.83 451.84 451.89
	451.9 453.1 453.2 453.3
	453.41 453.42 453.51 453.52
	453.6 453.71 453.72 453.73
	453.74 453.75 453.76 453.77
	453.79 453.8 453.81 453.82
	453.83 453.84 453.85 453.86

unspecified site
452 Portal vein thrombosis
453 Other venous embolism and thrombosis
  453.1 Thrombophlebitis migrans
  453.2 Other venous embolism and thrombosis of inferior vena cava
  453.3 Other venous embolism and thrombosis of renal vein
  453.4 Acute venous embolism and thrombosis of deep vessels of lower extremity
  453.5 Chronic venous embolism and thrombosis of deep vessels of lower extremity
  453.6 Venous embolism and thrombosis of superficial vessels of lower extremity
  453.7 Chronic venous embolism and thrombosis of other specified vessels
  453.8 Acute venous embolism and thrombosis of other specified veins
  453.9 Other venous embolism and thrombosis of unspecified site
454 Varicose veins of lower extremities
455 Hemorrhoids
456 Varicose veins of other sites
457 Noninfectious disorders of lymphatic channels
458 Hypotension
459 Other disorders of circulatory system
  459.1 Postphlebitic syndrome
  459.3 Chronic venous hypertension (idiopathic)
  459.8 Other specified disorders of circulatory system
  459.9 Unspecified circulatory system disorder
	453.87 453.89 453.9 459.11 459.12 459.19 459.31 459.32 459.81 459.89 459.9

Heart Valve Replacement V43.3 Heart valve	V43.3

The reported CPT Codes in the table "indication" were also examined against the Current Procedural Terminology (CPT) to test their validity (i.e., in acceptable range). The following table (Table 4.10) shows the reported the CPT codes.

Table 4.10. CPT Codes for indications of the AHC Base Database's Subjects.

CPT Code Class	CPT Codes
Orthopedic Procedures	77.65 77.66 77.67 77.69 77.75 77.76 77.77 77.79 77.85 77.86 77.87 77.96 78.05 78.06 78.07 78.09 78.15 78.16 78.17 78.19 78.29 78.45 78.46 78.47 78.49 78.55 78.56 78.57 78.59 79.05 79.06 79.09 79.15 79.16 79.19 79.25 79.26 79.29 79.35 79.36 79.39 79.56 79.65 79.66 79.75 79.76 79.85 79.86 79.89 80.25 80.26 80.27 80.6 80.75 80.76 81.44 81.45 81.46 81.47 81.49 81.51 81.52 81.53 81.54 81.55 81.56 81.57 81.59 81.62 81.63 81.64 81.65 81.66 84.11 84.12 84.13 84.14 84.15 84.16 84.17 84.18 84.19 84.51 84.58 84.59 84.61 84.62 84.65 84.81 84.84

On average, each subject had 1.83 indication record (range 1-14). The following table (Table 4.11) depicts the distribution of number of indication records per subject.

Table 4.11. Distribution of number of indication records among the subjects with indication records.

Number of Indication Records/Subject	Number of Subjects (%)
1	50196 (50.14)
2	29253(29.22)
3	12701(12.68)
4	4990(4.98)
5	1885(1.88)
6	688(0.68)

7	251(0.25)
8	97(<0.001)
9	23(<0.001)
10	9(<0.001)
11	4(<0.001)
12	1(<0.001)
13	4(<0.001)
14	1(<0.001)
Total	100103 (100%)

18.46% of the indication records were reported before patients' first

prescription start date of an anticoagulation medication at AHC. The

distribution of the reported days of indications are depicted in Table 4.12.

Table 4.12. Distribution of day numbers in which the indications were
recorded before or after patients' start date of anticoagulation therapy.

Recorded Day of Indication	Number of Indication Records (%)
Within 12 weeks before the start date	13768(7.51)
Within 4 weeks before the start date	11875(6.47)
Within 1 week before the start date	9426(5.14)
On the start date	26429(14.42)
Within 1 week after the start date	35628(19.44)
Within 4 weeks after the start date	44986(24.54)
Within 12 weeks after the start date	54856(29.93)

The complete duplicated records and partially duplicated ones (i.e., records

with the same indication on different days) were identified and excluded

(17.58% of the records). The following table (Table 4.13) depicts the

distribution of number of unique indications among the subjects with indication

records.

Table 4.13. Distribution of number of unique indications among the subjects
with indication records.

Number of Unique Indications/Subject	Number of Subjects (%)
1	64360 (64.29)

2	25330(25.30)
3	7056(7.04)
4	2227(2.22)
5	730(0.72)
6	267(0.26)
7	94(<0.001)
8	23(<0.001)
9	5(<0.001)
10	4(<0.001)
11	5(<0.001)
12	1(<0.001)
13	1(<0.001)
Total	100103 (100%)

- Comorbidity: In the AHC Base Database's table "comorbidity", each subject

has multiple comorbidity records. Each comorbidity record contained the

attributes {SURROGATE ID, DAY, CODE, SOURCE_SYSTEM}. As a quality

measure, first, the reported codes were examined against the published ICD-

9-CM 2011 codes (ICD-9-CM, 2014) to make sure that they were in the

acceptable range of the codes. All of the reported codes were in acceptable

range (Table 4.14).

Table 4.14. ICD-9 Codes for comorbidities of the AHC Base Database's Subjects.

ICD-9-CM 2011 Diagnostic and Procedure Code Class	Comorbidity codes reported at the AHC Base Database

Myocardial Infarction
  410 Acute myocardial infarction
  410.0 Of anterolateral wall
  410.1 Of other anterior wall
  410.2 Of inferolateral wall
  410.3 Of inferoposterior wall
  410.4 Of other inferior wall
  410.5 Of other lateral wall
  410.6 True posterior wall infarction
  410.7 Subendocardial infarction
  410.8 Of other specified sites
  410.9 Unspecified site
	410.01 410.02 410.11 410.12
	410.21 410.22 410.31 410.32
	410.41 410.42 410.51 410.52
	410.61 410.62 410.71 410.72
	410.81 410.82 410.91 410.92

Pulmonary Embolism
  415 Acute pulmonary heart disease
  415.1 Pulmonary embolism and infarction
	415.11 415.12 415.13 415.19

Cerebrovascular Disease
  431 Intracerebral hemorrhage
  432 Other and unspecified intracranial hemorrhage
    432.1 Subdural hemorrhage
  433 Occlusion and stenosis of precerebral arteries
    433.1 Occlusion and stenosis of carotid artery
    433.2 Occlusion and stenosis of vertebral artery
    433.8 Occlusion and stenosis of other specified precerebral artery
    433.9 Occlusion and stenosis of unspecified precerebral artery
  434 Occlusion of cerebral arteries
    434.1 Cerebral embolism
    434.9 Cerebral artery occlusion unspecified
  435 Transient cerebral ischemia
    435.1 Vertebral artery syndrome convert
    435.2 Subclavian steal syndrome convert
    435.3 Vertebrobasilar artery syndrome convert
    435.8 Other specified transient cerebral ischemias convert
    435.9 Unspecified transient cerebral ischemia convert
  436 Acute, but ill-defined, cerebrovascular disease
  437 Other and ill-defined cerebrovascular disease
    437.1 Other generalized ischemic cerebrovascular disease convert
    437.9 Unspecified cerebrovascular disease convert
  438 Late effects of cerebrovascular disease
	431 432.1 432.9 433.01 433.11 433.21 433.31 433.81 433.91 434.01 434.11 434.91 435.1 435.2 435.3 435.8 435.9 436 437.1 437.9 438.11 438.12 438.13 438.14 438.19 438.21 438.22 438.31 438.32 438.41 438.42 438.51 438.52 438.53 438.6 438.7 438.81 438.82 438.83 438.84 438.85 438.89 438.9

438.0 Late effects of cerebrovascular disease, cognitive deficits 438.1 Speech and language deficits 438.2 Hemiplegia/hemiparesis 438.3 Monoplegia of upper limb 438.4 Monoplegia of lower limb 438.5 Other paralytic syndrome 438.6 Late effects of cerebrovascular disease, alterations of sensations 438.7 Late effects of cerebrovascular disease, disturbances of vision 438.8 Other late effects of cerebrovascular disease 438.9 Unspecified late effects of cerebrovascular disease	
Gastrointestinal Hemorrhage 578 Gastrointestinal hemorrhage 578.9 Hemorrhage of gastrointestinal tract, unspecified	578.9
Hemorrhage in Head and Neck 784 Symptoms involving head and neck 784.7 Epistaxis 784.8 Hemorrhage from throat	784.7 784.8
Hemorrhage in Respiratory System 786 Symptoms involving respiratory system and other chest symptoms 786.3 Hemoptysis	786.3
Phlebitis and Thrombophlebitis 451 Phlebitis and thrombophlebitis 451.1 Phlebitis and thrombophlebitis of deep veins of lower extremities 451.2 Phlebitis and thrombophlebitis of lower extremities, unspecified 451.8 Phlebitis and thrombophlebitis of other sites 451.9 Phlebitis and thrombophlebitis of unspecified site 452 Portal vein thrombosis	451.11 451.19 451.2 451.81 451.82 451.83 451.84 451.89 451.9 453.1 453.2 453.3 453.41 453.42 453.51 453.52 453.6 453.71 453.72 453.73 453.74 453.75 453.76 453.77 453.79 453.8 453.81 453.82 453.83 453.84 453.85 453.86 453.87 453.89 453.9

453 Other venous embolism and thrombosis
  453.1 Thrombophlebitis migrans
  453.2 Other venous embolism and thrombosis of inferior vena cava
  453.3 Other venous embolism and thrombosis of renal vein
  453.4 Acute venous embolism and thrombosis of deep vessels of lower extremity
  453.5 Chronic venous embolism and thrombosis of deep vessels of lower extremity
  453.6 Venous embolism and thrombosis of superficial vessels of lower extremity
  453.7 Chronic venous embolism and thrombosis of other specified vessels
  453.8 Acute venous embolism and thrombosis of other specified veins
  453.9 Other venous embolism and thrombosis of unspecified site
454 Varicose veins of lower extremities
455 Hemorrhoids
456 Varicose veins of other sites
457 Noninfectious disorders of lymphatic channels
458 Hypotension
459 Other disorders of circulatory system
  459.1 Postphlebitic syndrome
  459.3 Chronic venous hypertension (idiopathic)
  459.8 Other specified disorders of circulatory system
  459.9 Unspecified circulatory system disorder	

About 1/3 (37.46%) of the AHC Base Database's subjects had comorbidity

records. On average within the subjects with comorbidity records, each

subject had 1.61 comorbidity record (range 1-13). The following table (Table

14.15) depicts the distribution of number of comorbidity records per subject.

Table 4.15. Distribution of number of comorbidity records among the subjects
with comorbidity records.

Number of Comorbidity Records/Subject	Number of Subjects (%)
1	36118(61.22)
2	14739(24.98)
3	5188(8.79)
4	1851(3.13)
5	685(1.16)
6	254(0.43)
7	109(<0.001)
8	29(<0.001)
9	7(<0.001)
10	2(<0.001)
11	4(<0.001)
12	1(<0.001)
13	1(<0.001)
Total	58988(100%)

18.36% of the comorbidity records were reported before patients' first

prescription start date of an anticoagulation medication at AHC. The

distribution of the reported days of comorbidities are depicted in table 4.16.

Table 4.16. Distribution of day numbers in which the comorbidities were
recorded before or after patients' start date of anticoagulation therapy.

Recorded Day of Comorbidity	Number of Comorbidity Records (%)
Within 12 weeks before the start date	7697(8.13)
Within 4 weeks before the start date	6588(6.95)
Within 1 week before the start date	5215(5.50)
On the start date	13971(14.75)
Within 1 week after the start date	20553(21.71)
Within 4 weeks after the start date	26782(28.29)
Within 12 weeks after the start date	32354(34.17)

Having compared the records of tables "indication" and "comorbidity", it was revealed that 69.26% of the comorbidity records were shared with indication records. Given this fact, the two quality controlled and cleaned tables were merged to create a new table called "morbidity". Each record of this new table has attributes {SURROGATE_ID, DAY, CODE, SOURCE_SYSTEM}. About 2/3 (65.68%) of the AHC Base Database's subjects had morbidity records. On average within the subjects with morbidity records, each subject had 1.71 morbidity records (range 1-13). Table 4.17 depicts the distribution of number of morbidity records per subject.

Table 4.17. Distribution of number of morbidity records among the subjects with morbidity records.

Number of Morbidity Records/Subject	Number of Subjects (%)
1	57116(55.22)
2	29131(28.16)
3	11346(10.97)
4	3756(3.63)
5	1326(1.28)
6	468(0.45)
7	181(0.17)
8	68(<0.001)
9	10(<0.001)
10	3(<0.001)
11	8(<0.001)
12	2(<0.001)
13	1(<0.001)
Total	103416(100%)

38.55% of the morbidity records were reported before patients' first prescription start date of an anticoagulation medication at AHC. The distribution of the reported days of morbidities are depicted in table 4.18.

Table 4.18. Distribution of day numbers in which the morbidities were recorded before or after patients' start date of anticoagulation therapy.

Recorded Day of Morbidity	Number of Morbidity Records (%)
Within 12 weeks before the start date	16430(15.88)
Within 4 weeks before the start date	14191(13.72)
Within 1 week before the start date	11340(10.96)
On the start date	29200(28.23)
Within 1 week after the start date	37351(36.11)
Within 4 weeks after the start date	44180(42.72)
Within 12 weeks after the start date	55412(53.58)

- Interacting Medications: In the AHC Base Database's table "interacting_medication", each subject has multiple medication records. Each interacting medication record contained the attributes {SURROGATE ID, DAY, MEDICATION_NAME, FREQ, DOSE_QTY, DOSE_QTY_UNIT, SOURCE_SYSTEM}. In the table, the entire columns of {DOSE_QTY, DOSE_QTY_UNIT} were completely blank. The column {FREQ} was partially complete. The cells under column {MEDICATION_NAME} were populated with free-text. Before conducting any cleaning or quality control measures, we had to extract information from the free-text in cells under column {MEDICATION_NAME} and translate them into a structured format. Through a text analysis step, the required data from the free text were extracted and presented in normalized and consistent structured formats. We used a multistep process which involved parsing the free texts into their components, normalizing the identified components, and extraction of the required data elements (i.e., name of medication, dosage, unit, and route of administration). The rate of incompleteness of the resulted records was high. Table 4.19 depicts the distribution of interacting medication records by each medication in the cleaned and controlled interacting_medication table and also presents

an example of the rate of completeness (i.e., percentage of each interacting medication's records with dosage information).

99.02% of the interacting medications records included one of the three medications of warfarin (37.85%), heparin (36.24%) and clopidogrel (24.93%). Warfarin records which have dosage information have the highest percentage of whole the records (11%). Although the number of heparin records was higher than the number of clopidogrel records, the percentage of clopidogrel records with dosage information out of whole the records was higher than that of heparin.

Table 4.19. Distribution of number of subjects under each interacting medication and the rate of medication records' dosage information completeness.

Interacting Medication	Number of Subjects under Interacting Medication	Rate of the Records' Dosage Information Completeness
Amiodarone	74,102	70.03%
Simvastatin	71,537	33.4%
Fluvastatin	61,517	68.03%
Lovastatin	1,793	89.18%
Atrovastatin	2,292	22.9%
Rosuvastatin	2886	71.58%
Pravastatin	434	73.5%
Aspirin	1,310	90.15%

The above table indicates that a big number of subjects have received more than one medication. The following table (Table 4.20) shows the number of subjects who have been under treatment with 1 to 6 medications.

Table 4.20. Distribution of subjects by number of medications.

No. of Medications	1	2	3	4	5	6
No of Subjects	75282	16441	3049	358	20	1

- INR: In the AHC Base Database, each subject has longitudinal INR records with multiple INR records for each measurement day which were not on a regular basis. Each INR record contained the attributes {"SURROGATE_ID", "EFFECTIVE_DAY", "LAB_DESCRIPTION", "LAB_VALUE", "SOURCE_SYSTEM"}. Since INR measurement can vary within a specific time frame due to factors such as different measurement methods (e.g., medical laboratory testing, point-of-care testing, patient self-testing), we created a cleaning algorithm for INR records. Before applying the algorithm, first, each INR record was examined to identify and exclude any INR record whose INR lab value was missing. Then the INR cleaning algorithm was applied (Figure 4.6).
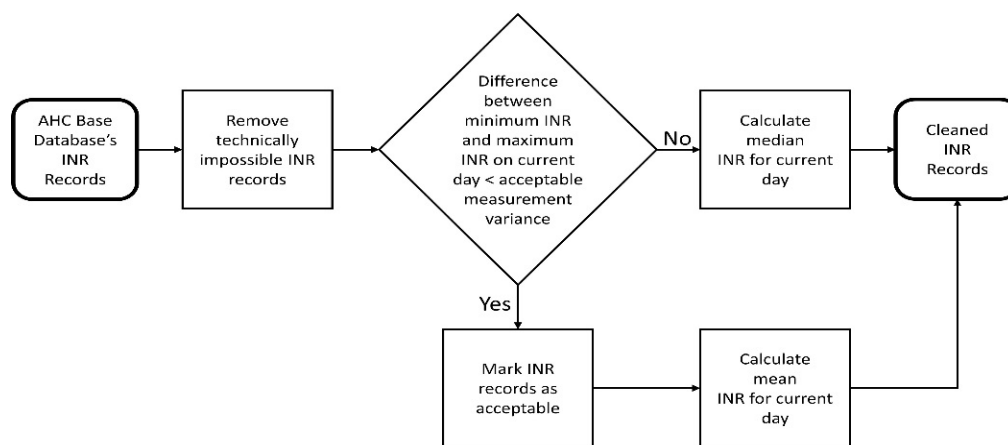
Figure 4.6. INR records cleaning algorithm.

According to the algorithm, we first removed all INR values we considered technically not acceptable according to the AHC's laboratory guidelines (i.e., $0.9 >$ INRs > 10). Then we took the next steps of the algorithm. If all INR

records for a single subject in a single day differed (i.e., difference between minimum INR and maximum INR for the same day) by less than acceptable measurement variance of 0.5 INR-units (Rasmussen, 2012), then all INR records for that subject for that day were marked acceptable by the algorithm and the mean of the INR records was calculated representing INR value for the day. For all other days, median INR for each day was calculated. The rate of INR record completeness for the AHC Base Database is 77.03%. The subjects with INR records had 19.17 INR records on average.

4.2.2.) WiAD Database and Data Load

The next step in the process of developing a research anticoagulation database was to load the quality controlled and transformed data into a working database now called "WiAD" standing for "Wisconsin Anticoagulation/Anti-clotting Database". The database was implemented in a MySQL Server and i2b2. Two specifics tasks were undertaken in this process: designing data models consistent with the research needs and also making some vocabularies and ontologies across the data models to make sure that the queries run against the implemented databases would consistently return the data meeting the research oriented criteria.

4.2.2.1.) MySQL Database

WiAD was implemented in MySQL Server as one of its database management systems. The database could be queried directly through the MySQL server or it could be done using a tool called WiAD-Miner (explained in section 4.3.4). The data model of WiAD on the MySQL server is depicted in Figure 4.7. In this star schema, table "patients" is the core table which

basically stores non-longitudinal attributes of the subjects (e.g., Gender, Age, and Race). The other tables store the subjects' longitudinal attributes.
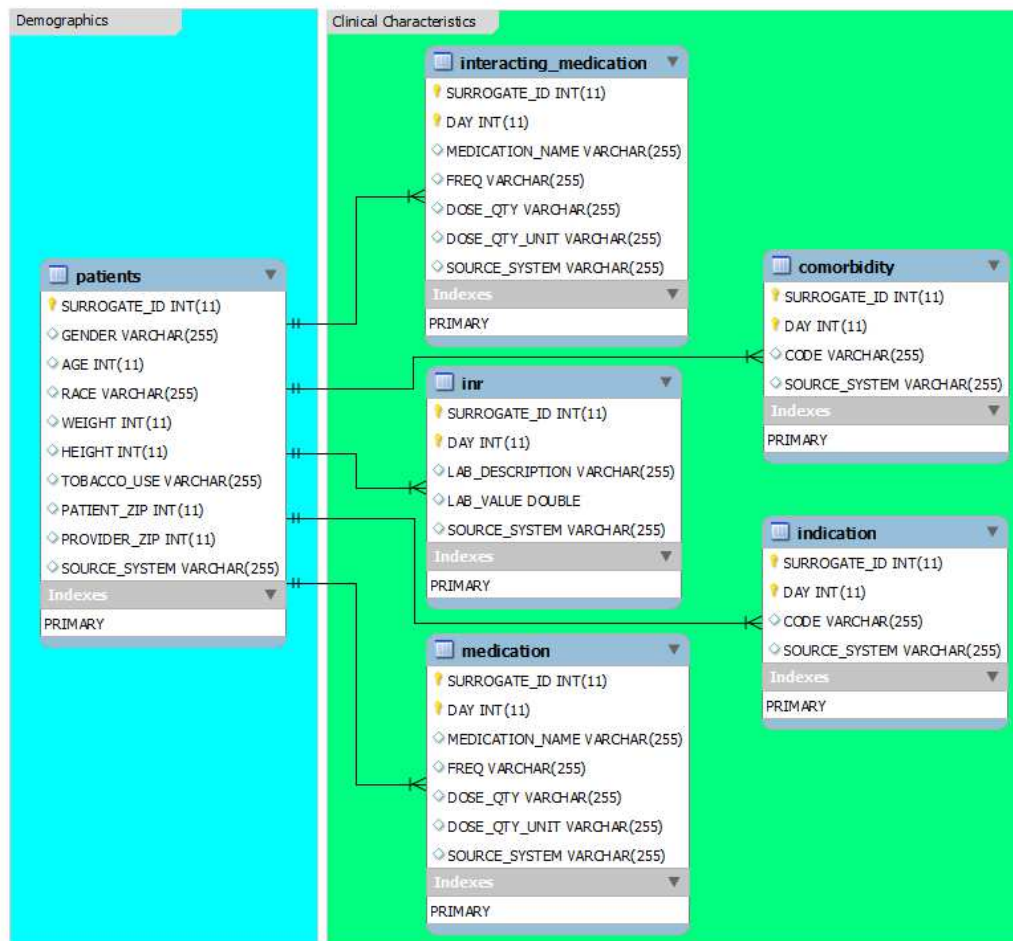
Figure 4.7. Entity Relationship (ER) Model of the WiAD's MySQL Database.

4.2.2.2.) i2b2 Database

Our anticoagulation/anticlotting research database "WiAD" was also implemented in i2b2. The data model provided in the i2b2 database is called "star schema" where tables are connected as a star. Figure 4.8 shows the star schema consisting of Observation Fact surrounded by Patient Dimension, Visit Dimension, Concept Dimension, and Provider Dimension for the WiAD database.

In this schema, the Observation Fact table represents a patient object and
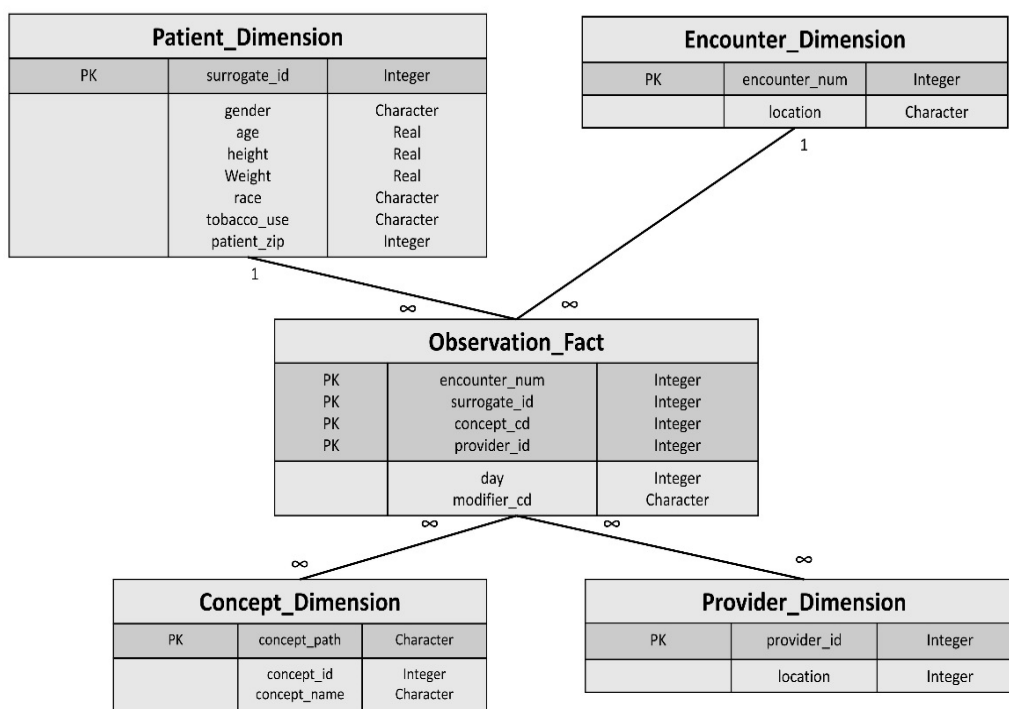
Figure 4.8. WiAD's i2b2 Data Model.

other four dimensions represent its attributes such as who (patient

information), when (dates), what (ontology for clinical patients' data) and

where (hospital or treatment facility), respectively.

Table 4.21. Definitions of the concepts of the WiAD's i2b2 data model.

Concept	Definition
surrogate_id	Patient unique identification number
encounter_num	Patient visit number (surrogate_id + day)
concept_cd	Code for observation of interest (i.e., morbidity, medications, lab test)
provider_id	Provider unique identification number
day	Starting date-time of observation (i.e. "DAY")
concept_path	Ontology path for concepts
concept_id	Unique identification number for concepts
concept_name	Actual name of the concept
location	Zipcode of the provider

In i2b2, ontology represents a data model of a target domain. Ontology is stored in Concept Dimension table, which contains a symbolic name of an individual attribute (Concept ID) and a path from a root of ontology to an individual attribute (Concept Path). Every attribute is allowed to have only one conceptual path, so that ontology in the i2b2 includes no multiple inheritance. The following table provides definitions for some of the specific concepts of the WiAD's i2b2 data model.

The following table (Table 4.22) details the WiAD's i2b2 ontology.

Table 4.22. Hierarchical structure of the WiAD's i2b2 ontology.

Demographics
Demographics\Age\
Demographics\Gender\
Demographics\Gender\Female\
Demographics\Gender\Male\
Demographics\Height\
Demographics\Patient Zipcode\
Demographics\Provider Zipcode\
Demographics\Race\
Demographics\Race\African American\
Demographics\Race\American Indian or Alaskan Native\
Demographics\Race\Asian\
Demographics\Race\Hispanic\
Demographics\Race\Not Asked\
Demographics\Race\Null\
Demographics\Race\Other\
Demographics\Race\Pacific Islander\
Demographics\Race\Unknown\
Demographics\Race\White\
Demographics\Tobacco Use\
Demographics\Weight\
Morbidity
Morbidity\Diagnosis\
Morbidity\Diagnosis\Atherosclerosis of Aorta\
Morbidity\Diagnosis\Atrial Fibrillation\
Morbidity\Diagnosis\Atrial Flutter\
Morbidity\Diagnosis\Bleeding\
Morbidity\Diagnosis\Coronary Artery Disease\
Morbidity\Diagnosis\Deep Vein Thrombosis\
Morbidity\Diagnosis\Diabetes Mellitus\
Morbidity\Diagnosis\Heart Failure\
Morbidity\Diagnosis\Hypertension\

Morbidity\Diagnosis\Myocardial Infarction\ Morbidity\Diagnosis\Peripheral Artery Diseases\ Morbidity\Diagnosis\Pulmonary Embolism\ Morbidity\Diagnosis\Stroke\ Morbidity\Procedure\ Morbidity\Procedure\Heart Valve Replacement\ Morbidity\Diagnosis\Orthopedic Procedures\
Medications Medications\Anticoagulation Agent\Abciximab Medications\Anticoagulation Agent\Clopidogrel\ Medications\Anticoagulation Agent\Dabigatran\ Medications\Anticoagulation Agent\Dipyridamole\ Medications\Anticoagulation Agent\Eptifibatide\ Medications\Anticoagulation Agent\Heparin\ Medications\Anticoagulation Agent\Ticlopidine\ Medications\Anticoagulation Agent\Tirofiban\ Medications\Anticoagulation Agent\Warfarin\ Medications\Interacting Medication\Amiodarone\ Medications\Interacting Medication\Aspirin\ Medications\Interacting Medication\Atrovastatin\ Medications\Interacting Medication\Fluvastatin\ Medications\Interacting Medication\Lovastatin\ Medications\Interacting Medication\Pravastatin\ Medications\Interacting Medication\Rosuvastatin\ Medications\Interacting Medication\Simvastatin\
Lab Test Labe Test\INR\

4.2.2.3.) Ontology issues

Usually in clinical datasets, there exist different data types in data which can each be supported by a separate dictionary such as the ones we have developed for the AHC dataset. The advantage of using ontologies and dictionaries is that they help standardize raw data and also add logical hierarchical structure. In the case of i2b2, these are essential component of its ontology too. In the ideal situation and with the aim of interoperability, the data are mapped to reference standard dictionaries and ontologies. With the use of reference standard dictionaries and ontologies (e.g., Drug Bank), we could subgroup data types then support data elements across all data sources.

Even though there are different reference dictionaries and ontologies to use for normalizing and standardizing data (e.g., ICD-9, CPT, LOINC, NDC), the standard ontologies used are largely based on the data available. For the AHC dataset, our diagnosis data are coded in ICD-9, so that is the hierarchy we use in our ontology. The same holds true for procedures. With our medication and laboratory data, it is not coded in terminology with a standard hierarchy, so we had to organize the data using other locally developed dictionaries.

4.2.3.) Data Analysis Methods and Tools

As introduced, WiAD is a longitudinal database which includes subjects' characteristics, treatment plans and outcomes that have been at multiple follow-up times. A longitudinal database generally provides multiple or repeated measurements and records on each subject. Accordingly and given that such repeated measurements and records are correlated within subjects, there are needs for special analysis and inference techniques for a longitudinal dataset. Although longitudinal data are very beneficial but there challenges in using such data which in some cases are not without cost. There are several challenges posed (Heagerty, 2014):

- Participants follow-up. There is always the possibility of bias as the retrospective longitudinal EMR data might include the data of subjects who might have incomplete follow-up or drop-out during their treatment periods. Accordingly, the analysis of such data requires special attention to make sure that the extracted subset of the data meets the assumptions of the study that it is going to be used for.

- Analysis of correlated data. Given the nature of the longitudinal of EMR data, we need to use statistical methods that can account for the intra-attributes and /or intra-subjects correlations.

- Time-varying covariates. Although longitudinal data provide the opportunity to study the association between the changes in one subject's attribute by changes in other attributes or outcome of interest, the direction of causality can be complicated by feedback between the attributes.

In the following section, some of the aspects of the longitudinal WiAD data is presented.

4.2.3.1.) Landscape of the WiAD Data

In the previous sections, it was explained the methods that we have applied to clean and quality control some of the longitudinal aspects of the WiAD data. However, there are some aspects of the data that should be managed and adjusted based on each study's goals and design. For instance, one of the potential cross-attribute studies in the WiAD could be to study the correlation between the changes in warfarin doses and the subjects' INR values.
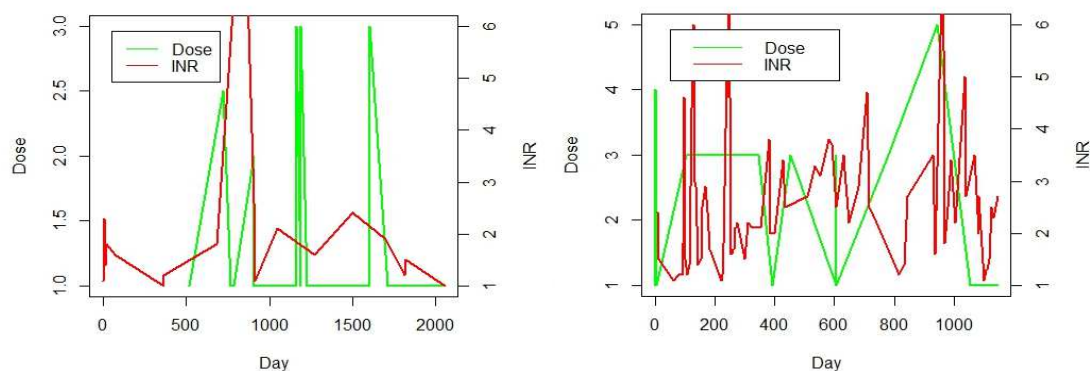
Figure 4.9.Warfarin dose changes versus INR values changes for two WiAD's subjects.

However, the fact is that longitudinal data patterns for these attributes are very complex and affected by different factors so that it requires applying some specific methods to extract the data records which are appropriate for such a study. Figure 4.9 demonstrates the warfarin recorded doses versus INR recorded values for two different WiAD's subjects.

Taking into consideration that INR is used to monitor different anticoagulation agents and also the fact that a given patient under anticoagulaton therapy might receive different types of anticoagulation agents during his/her treatment period depending on his/her treatment condition (e.g., ambulatory, in-patient care or surgical procedure), it is important to find the desired warfarin exposure periods in the EMR data and extract the associated data. Given this complexity, we have developed some algorithms which help extract data from WiAD based on desired longitudinal patterns such as treatment periods (e.g. fixed time or medical procedure periods), treatment indications which require different target therapeutic INR ranges, and interval and frequency between successive observed dose records or INR values. These features have been implemented in WiAD-Miner explained in the next section.

4.2.3.2.) WiAD-Miner

An interactive data profiling and population "segmentation" tool for WiAD (WiAD-Miner) has been developed in R (R Core Team, 2014) using RStudio (RStudio, 2014). WiAD-Miner (Figure 4.10) which also has a web application version includes all the profiling, outcome metric, and related data analysis functions for anticoagulation agents. WiAD-Miner includes a cohort selection tool to profile and identify patient subpopulations by any or a combination of patients' characteristics including gender, age group, race, patient's residence

zipcode, provider's zipcode, medication, medication exposure, duration of treatment, number of dose records, frequency of INR values, medication indication, and comorbidities. WiAD-Miner presents a clear view of each
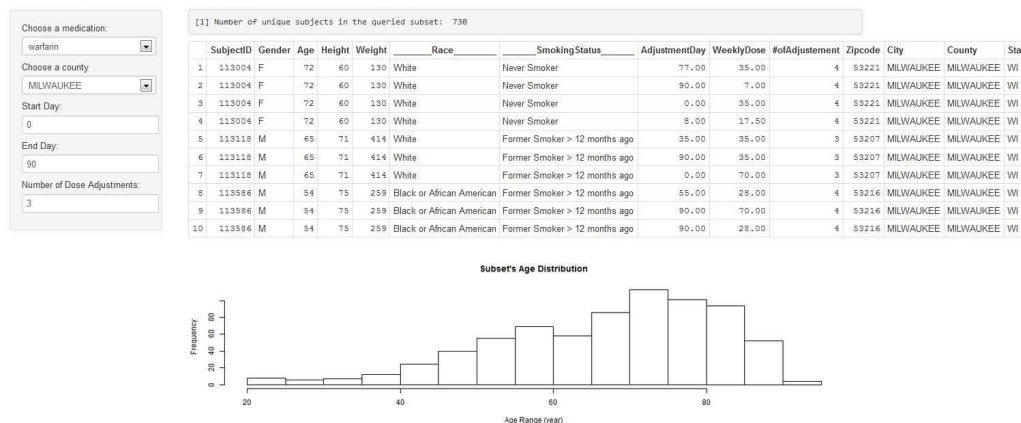
Figure 4.10. A screenshot of the WiAD-Miner's interface.

extracted subset by producing statistical characteristics and visual profiling and allows adjustment of various parameters such as the medication exposure period definition, triggering a re-profiling and thereafter, re-calculation of outcome metrics.

4.2.4.) Estimating Parameters for PK/PD

Our *in silico* simulation model takes advantage of Hamberg's PK/PD model (Hamberg, 2007) for individualization of warfarin therapy. As introduced in Chapter 3, Hamberg et al. have characterized the relationship between warfarin dose and international normalized ratio (INR) response and they have identified CYP2C9 genotype and age as predictors for S-warfarin clearance, and VKORC1 genotype as a predictor for warfarin sensitivity. Our *in silico* platform is able to take advantage of domain knowledge and integrate population characteristics such as genotype distributions if they are not provided for the study populations. Accordingly and given that WiAD originally

did not include genotype information, we have derived distribution of required

genotypes for our study populations from some studies such as the one

conducted by Scott (2010). A recent study by Scott (2010) has provided the

allele frequencies of some of the principal genes known to influence

interindividual warfarin dose variability (e.g., CYP2C9 and VKORC1) in

African-American, Asian and White populations. We have used these

information to impose population genotype distribution information into our *in

silico* platform. Table 4.23 summarizes the CYP2C9 allele and genotype

frequencies.

Table 4.23. CYP2C9 Genotype Frequencies (Scott, 2010).

CYP2C9	African-American	Asian	White
Extensive Metabolizer			
*1/*1	75.5	86.3	66
Intermediate Metabolizer			
*1/*2	4.3	3.9	15.1
*1/*3	3.3	6.9	9.4
*1/*5	2.7	0	0
*1/*6	1.7	0	0
*1/*8	8.7	1	0
*1/*11	2	0	0.9
Poor Metabolizer			
*2/*2	0.3	1	6.6
*2/*3	0.3	0	1.9
*2/*8	0	0	0
*3/*3	0	0	0
*3/*5	0	0	0
*3/*8	0	1	0
*3/*11	0.3	0	0
*5/*6	0.3	0	0
*8/*11	0.3	0	0

Table 4.24 demonstrates the VKORC1G 1639G>A allele frequencies.

Table 4.24. VKORC1 Genotype Frequencies (Scott, 2010).

VKORC1 1639G>A	African-American (%)	Asian (%)	White (%)
G/G	80.3	22.5	36.8
G/A	17.7	21.6	45.3
A/A	2	55.9	17.9

Table 4.25 summarizes the combined CYP2C9 and VKORC1 genotype

frequencies for some of the study populations such as African-American,

Asian and White. As depicted in the table, a majority of White and Asian

individuals carry a variant of CYP2C9 and VKORC1 compared with African-

American.

Table 4.25. Combined CYP2C9 and VKORC1 Genotype Frequencies (Scott, 2010).

CYP2C9	VKORC1 1639G>A	African-American (%)	Asian (%)	White (%)
Extensive Metabolizer	G/G	63.4	20.6	41.5
	G/A	0	0	0
	A/A	1.4	55	13.3
Intermediate Metabolizer	G/G	20.3	7.9	10.3
	G/A	2.9	3	10.4
	A/A	0.7	1	4.7
Poor Metabolizer	G/G	0	2	4.7
	G/A	0.4	0	3.7
	A/A	0	0	0

4.3.) Results

This effort has resulted in an EMR-based longitudinal

anticoagulation/anticlotting database, WiAD, using Aurora Health Care's

electronic medical records. This database is a seminal translational

anticoagulation research tool that support US-prioritized "Secondary Use" of

electronic medical records to improve health care and comparative effectiveness research studies. The database could be used to develop and derive conditional parametric models for WiAD-based studies such as studies that are presented in Chapters 5 and 6.

4.4.) Conclusion and Future Work

This effort has shown that EMR longitudinal data is a rich resource to develop research-grade databases. Our effort is aligned with current national developments such as PCORnet aiming to leverage the secondary use of EMR data for research purposes. As presented in the next chapters, WiAD database has significant potential for comparative effectiveness research and conducting patient-centered outcomes research. In future, we plan to improve the database by including more cardiovascular associated patient characteristics and also embedding more robust data transforming and extraction algorithms.

4.5.) References:

Bayley, K. B., Belnap, T., Savitz, L., Masica, A. L., Shah, N., & Fleming, N. S. (2013). Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. Medical care, 51(8 Suppl 3), S80-86.

Centers for Disease Control and Prevention (CDC). (2014). International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Retrieved June15, 2014, from http://www.cdc.gov/nchs/icd/icd9cm.htm

Clinical Data Research Networks(CDRNs). (2014). Patient-Centered Outcomes Research Institute. Retrieved September 10, 2014 from http://www.pcori.org/content/clinical-data-and-patient-powered-research-networks-awarded-projects

Cruz-Correia, R. J., Rodrigues, P., Freitas, A., Almeida, F. C., Chen, R., & Costa-Pereira, A. (2010). Data quality and integration issues in electronic health records. Information discovery on electronic health records, 55-95.

Elkin, P. L., Trusko, B. E., Koppel, R., Speroff, T., Mohrer, D., Sakji, S., et al. (2010). Secondary use of clinical data. Studies in health technology and informatics, 155, 14-29.

Heagerty P.J. Longitudinal Data Analysis. Retrieved October 9, 2014, from http://faculty.washington.edu/heagerty/ 2014

Hristidis, V. (2010). Information discovery on electronic health records: CRC Press.

ICD9-11-CM. (2014). Centers for Disease Control and Prevention. Retrieved October 20, 2014 from http://www.cdc.gov/nchs/icd/icd9cm_maintenance.htm

Maskin, L. P., Attie, S., Setten, M., Rodriguez, P. O., Bonelli, I., Stryjewski, M. E., et al. (2010). Accuracy of weight and height estimation in an intensive care unit. Anaesthesia and intensive care, 38(5), 930-934.

Muthalagu, A., Pacheco, J. A., Aufox, S., Peissig, P. L., Fuehrer, J. T., Tromp, G., et al. (2014). A rigorous algorithm to detect and clean inaccurate adult height records within EHR systems. Applied clinical informatics, 5(1), 118-126.

Prokosch, H. U., & Ganslandt, T. (2009). Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods of information in medicine, 48(1), 38-44.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved August 1, 2014, from http://www.R-project.org/

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

Rasmussen, R. S., Corell, P., Madsen, P., & Overgaard, K. (2012). Effects of computer-assisted oral anticoagulant therapy. Thrombosis journal, 10(1), 17.

RStudio. (2014). RStudio: Integrated development environment for R (Version 0.98.1091) [Computer software]. Boston, MA. Retrieved November 6, 2014 from http://www.rstudio.com/

Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., et al. (2007). Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association: JAMIA, 14(1), 1-9.

Scott, S. A., Khasawneh, R., Peter, I., Kornreich, R., & Desnick, R. J. (2010). Combined CYP2C9, VKORC1 and CYP4F2 frequencies among racial and ethnic groups. Pharmacogenomics, 11(6), 781-791.

Selby, J. V., Beal, A. C., & Frank, L. (2012). The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda. JAMA : the journal of the American Medical Association, 307(15), 1583-1584.

U.S. Census Bureau. Statistical Abstract of the United States: 2012. (2014). Retrieved July 16, 2014 from https://www.census.gov/compendia/statab/2012/tables/12s0209.pdf

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association : JAMIA, 20(1), 144-151.

Chapter 5: *In Silico* Translational Research Pharmacogenetic Comparative Effectiveness Studies

5.1.) Introduction

We have developed an *in silico* pharmacogenetic PC-CER framework for anticoagulation therapy. The framework includes options to simulate patient populations ("Clinical Avatars"), multiple initial dosing protocols including PG-based and non PG-based, multiple dose adjustment and maintenance protocols, PK/PD modeling and prediction of various types of outcome measures. We have validated the framework against two major warfarin clinical trials, CoumaGen-I and CoumaGen-II (Anderson, 2007a, 2012b). Then, we have used our highly adaptable *in silico* framework to conduct 90-day anticoagulation therapy simulations for the Aurora Heath Care (AHC) state-wide warfarin patient population using a collection of warfarin dosing protocols to study comparative effectiveness and to identify different optimal protocols depending on subpopulations defined by patient characteristics.

5.2.) CoumaGen-I Clinical Trial Simulation

5.2.1.) Background and Objective

The details of this study were published by Fusaro (2013) in the journal Circulation. In this study, the accuracy of the clinical trial simulator of our *in silico* PC-CER framework was demonstrated by reproducing the CoumaGen-I clinical trial outcomes ("CoumaGen-I Simulation 1"), and then the simulator was used to evaluate a new dosing protocol ("CoumaGen-I Simulation 2"), to determine whether this new study design was significantly more beneficial for the same population.

5.2.2.) Methods

5.2.2.1) Study Design

- The original study. The original CoumaGen-I clinical trial study (Anderson, 2007) was designed as a prospective, randomized study comparing PG-based and standard empirical dosing in patients being initiated on oral anticoagulation. The study objectives were prospectively to validate a PG-based dosing algorithm (Anderson, 2007) and to assess its impact on INR-based efficacy and safety end points.

In the original study, the inclusion criteria to recruit subjects were as follows:
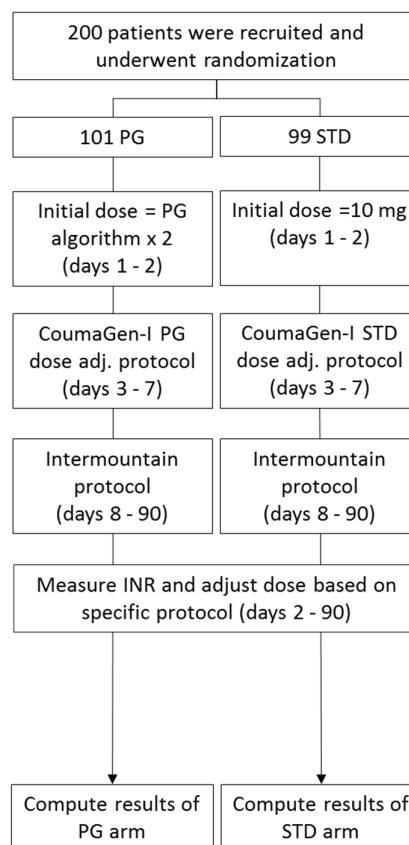
Figure 5.1. Study design for the original CoumaGen-I clinical trial. PG: Pharmacogenetic arm, STD: Standard arm.

age ≥18 years and an indication for anticoagulation with a target INR of 2 to 3. And the subjects with the following characteristics were excluded: women who were pregnant, lactating, or of child-bearing potential, those taking rifampin within 3 weeks, or patients with comorbidities precluding standard dosing (e.g., advanced physiological age, renal insufficiency/creatinine >2.5 mg/dL, hepatic insufficiency, terminal disease). Then, the 200 qualified subjects underwent blind randomization to the pharmacogenetic (PG) or standard (STD) arm. The study recruited 101 patients into the PG arm and 99 patients into the standard clinical STD arm. Figure 5.1 illustrates the study design of the original clinical trial.

- The simulation study. In this study, the two following simulations were designed and conducted: "CoumaGen-I Simulation 1" and "CoumaGen-I Simulation 2". For each of these, with the use of a Bayesian Network Model (BNM), a sufficient number of clinical avatars (n=200000) was created to conduct 1000 simulations. For each of the 1000 simulations, 101 avatars for the PG arm and 99 for the STD arm were randomly recruited from this large clinical avatar population. Then, following the specific dosing protocol for each arm explained in the following sections, daily dose and INR for each avatar for that arm for 90 days were predicted.

CoumaGen-I Simulation 1 followed the dosing protocol as specified in the original CoumaGen-I clinical trial (Anderson, 2007; Figure 5.1). The standard arm dosing followed the 10-mg warfarin nomogram from Kovacs (2003) for days 1 and 2 followed by dose adjustment based on INR according to the Kovacs (2003) protocol for days 3 to 7. For days 8 to 90, CoumaGen-I used the Intermountain Healthcare warfarin dosing algorithm. The CoumaGen-I

pharmacogenetic arm dosing followed a dosing algorithm which required both clinical and genotype (i.e., CYP2C9 and VKORC1) information to calculate doses for days 1 and 2. Like the original study, the initial doses were followed by a dose adjustment based on INR by multiplying standard arm changes by a pharmacogenetic algorithm coefficient for days 3 to 7. The pharmacogenetic algorithm coefficient was defined as the ratio of the estimated individual weekly dose determined by the pharmacogenetic algorithm to the standard weekly dose of 35 mg. For days 8 to 90, CoumaGen-I used the Intermountain Healthcare warfarin dosing protocol.
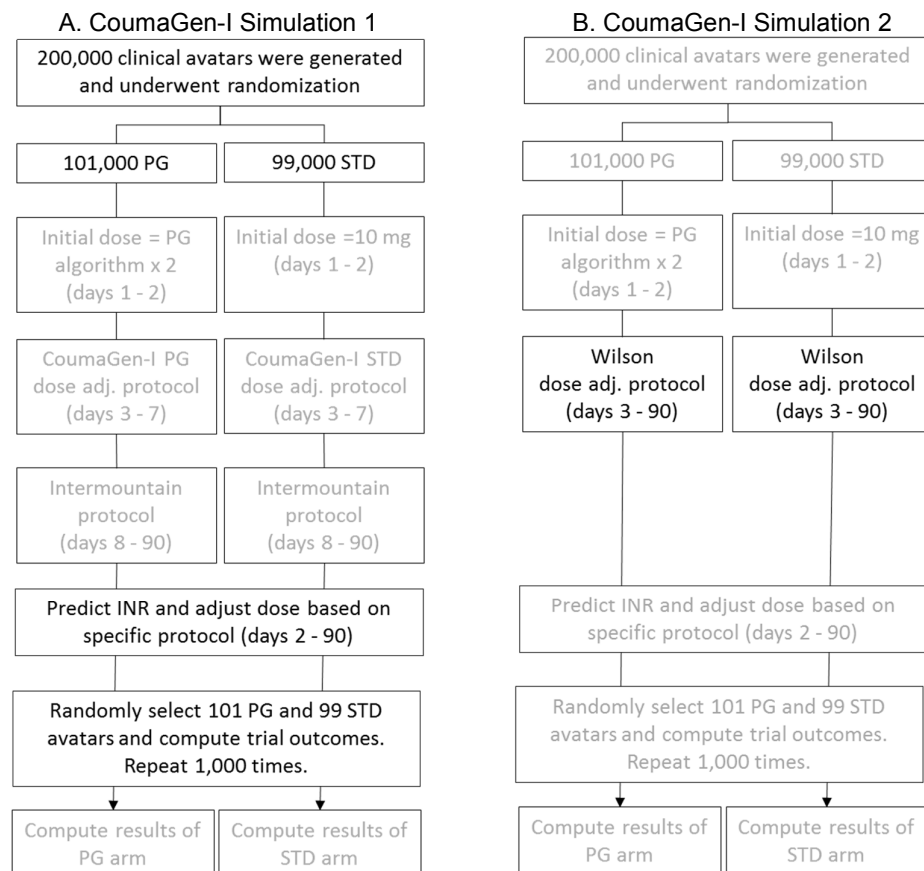
Figure 5.2. The study design of the CoumaGen-I Simulation 1 and 2. Black text represents new features compared to the original CoumaGen-I study, whereas gray text represents those features in common with the original study as depicted in Figure 5.1. PG: Pharmacogenetic arm, STD: Standard arm.

The design of the CoumaGen-I Simulation 2 was different from the original clinical trial. It was conducted using a third dosing protocol (nomogram) offered by Wilson (2007). In this study, the same starting doses were used for days 1 to 2 as in the original CoumaGen-I trial, 10-mg/d for the standard arm and 2 times the pharmacogenetic dose for the pharmacogenetic arm. For days 3 to 90, the Wilson protocol was used to increase or decrease the dose proportionally based on low or high INR values, respectively. The following figure (Figure 5.2) illustrates the design of the two simulation studies.

5.2.2.2) Clinical Avatar Populations

With regard to the BNM used for creating clinical avatars in this study, some consideration were made as the original CoumaGen-I's individual patients' information were not available. Taking into consideration the subjects' characteristics required for the dosing algorithms used in this study, the prior probabilities used for the study's BNM were based on the statistical characterization of the patient population including age, sex, weight, height, race, body surface area (calculated from height and weight), smoking status, deep vein thrombosis status, amiodarone use status, and genotypes for CYP2C9*2, CYP2C*3, and VKORC1. The CoumaGen-I trial data used age, sex, weight, height, deep vein thrombosis, smoking status, and genotypes as their variables. In the simulation study, some other required prior probabilities (e.g., amiodarone status) for the BNM were estimated by using data from the Centers for Disease Control and Prevention and the 2000 US Census. The extracted prior probabilities were mostly for whites as about 95% of the original CoumaGen-I trial's subjects were white.

Another challenge in developing a BNM for this study which helped produce physiologically realistic clinical avatars was the lack of joint conditional probability distributions for the characteristics. To deal with this case, some external data and information sources were used such as the US Census 2007 to 2008 Table 209 (http://www.

census.gov/compendia/statab/2012/tables/12s0210.pdf) which details height and weight distributions as functions of age and sex. The extracted normal distributions from the above sources were transformed to match the actual population in the CoumaGen-I trial. For instance, the distributions for subjects 40 to 49, 50 to 59, and 60 to 69 years of age were z-transformed and scaled according to the mean and standard deviation for the CoumaGen-I trial pharmacogenetic and standard arms, respectively. For use in the BNM, then a dependency table by sampling from these distributions and calculating the percentages for each age/sex group was developed. The BNM was implemented in TETRAD IV (Scheines, 1998) to produce the clinical avatar populations for the both Simulations 1 and 2.

5.2.2.3.) Measuring Outcome Metrics

The primary outcome metric was Percent Time INR in Therapeutic Range ("TTR"). Consistent with the original CoumaGen-I study, TTR was defined as the percentage of time an individual avatar had an INR between 1.8 and 3.2 during the 90-day simulation. Although our simulator could calculate daily INR values, only INRs on those days in which the INR would have been checked in the clinic according to the specific protocol were considered.

5.2.2.4.) Clinical Trial Simulations

As stated in previous sections, in the original CoumaGen-I randomized controlled trial, the 200 subjects were randomly assigned to two PG (n=101) and STD (n=99) arms. Accordingly, in the Simulation 1 and 2 studies, the 200,000 clinical avatars were also randomly sampled to recruit 101 and 99 avatars for the PG and the STD simulation arms. The number of created avatars provided the opportunity to create 1000 parallel arms for simulations. Then daily dose and INR for each avatar were predicted following a specific protocol for that arm for 90 days. The predicted doses, INR, INR-monitoring frequency, and population statistics for each clinical trial simulation and across all 1000 simulations were recorded to calculate and produce the mean, standard deviation, and probability value (unpaired t test) for TTR for each study arm and for the aggregated populations from the 1000 clinical trial simulations. The simulations produced predictions were then compared with the CoumaGen-I results. The simulations were implemented in R (R Core Team, 2013) and run on the affiliated institute's high performance computing environment.

5.2.3.) Results

Developing a BNM and then generating clinical avatars were one of the first steps taken in this study. Table 5.1 demonstrates the characteristics of the clinical avatar populations generated for both PG and STD simulation arms versus the characteristics of the PG and STD arms' populations in the original CoumaGen-I study.

A few tests were done to show that the clinical avatar populations generated based on the BNM were statistically similar to the original CoumaGen-I

population. As depicted in Table 5.1, the distribution of the characteristics across the PG arms and STD arm are very comparable and statistical test shows no significant difference (P >0.05). The variable dependency embedded in the BNM of the populations were also tested to see if they also persisted in the simulated clinical avatar populations. To do so, a log-linear model was fitted to all relevant associations and probability values by using the Pearson χ^2 statistic were calculated. The test indicated that there were no significant differences (P >0.05) between the dependencies in the clinical avatar populations and the original CoumaGen-I populations (Table 5.1). In addition, it was tested to make sure that the non-embedded dependencies in the BNM did not exist in the clinical avatar populations.

Table 5.1. Characteristics of the subjects in the original CoumaGen-I study versus that of Clinical Avatars generated for both PG and STD arms. DVT: deep vein thrombosis; PG: pharmacogenetic arm, SD: standard deviation, STD: standard clinical arm (Fusaro, 2013).

Characteristic	CoumaGen-I Original		Clinical Avatars	
	PG Arm	STD Arm	PG Arm	STD Arm
n	101	99	101	99
Age, y (mean)	63.2	58.9	62.5	58.3
Male, %	49.5	56.6	51.3	55.1
Weight, kg, mean±SD	92.1±24.6	94.7±24.2	89.8±24.3	91.9±24.3
DVT, %	18.8	28.3	18.6	28.6
White, %	94.1	94.9	94	95
CYP2C9*2, %				
CC	82	76.5	82.2	76.7
CT	18	23.5	17.8	23.3
TT	0	0	0	0
CYP2C9*3, %				
AA	89	87.6	88.9	87.9
AC	10	11.3	10.2	11.1
CC	1	1	1	1
VKORC1 1173, %				
G/G	50.5	34.7	50.5	34.6
G/A	35.4	50	35.6	50.2
A/A	14.1	15.3	13.9	15.2

The accuracy of the clinical trial simulator of our *in silico* PC-CER framework was validated by simulating the original CoumaGen-I clinical trial through conducting CoumaGen-I Simulation 1 (Figure 5.1 and 5.2A).

The CoumaGen-I Simulation 1 reproduced the primary TTR outcome of the original CoumaGen-I trial (Table 5.2). It predicted a mean TTR of 70.5% and 72.0% in the STD and PG arms, respectively. Similar to the original CoumaGen-I trial, the difference between the mean TTRs was not significant (P >0.05). The results of the CoumaGen-I Simulation 1 showed no statistical difference between predicted and actual TTRs for CYP2C9 extensive metabolizer, intermediate metabolizer, and poor metabolizer subsets (P >0.05). Although the original CoumaGen-I showed a significant 9.8% reduction in out-of-range INRs for the wild type and multiple variants subgroup in the PG arm, CoumaGen-I Simulation 1 indicated a nonsignificant 2.7% reduction in out-of-range INRs for the similar groups.

Through the second simulation, CoumaGen-I Simulation 2, it was tested if a modification to PG and STD arms of the original CoumaGen-I dosing algorithms would result in a significant change in outcomes of the two arms. As depicted in Figure 5.2B, all model and simulation components of the CoumaGen-I Simulation 1 (i.e., clinical avatars, initial dosing, PK/PD parameters, and TTR outcome calculations) remained the same with the exception of the replacement of part of the original CoumaGen-I dosing protocol with the Wilson protocol (Wilson, 2007) for days 3 to 90. The simulation, CoumaGen-I Simulation 2, was run and the outcome metrics were calculated as in the Simulation 1 (Table 5.2).

Table 5.2. Percent Time INR in Therapeutic Range (TTR) for some
subpopulations in 1000 simulated trials of CoumaGen-I Simulations 1 and 2.

Subpopulation	CoumaGen-I Simulation 1		CoumaGen-I Simulation 2	
	PG Arm	STD Arm	PG Arm	STD Arm
All Avatars, %, mean±SD	72±26.6	70.5±26.8	78.8±11.9	73.7±13.6
CYP2C9, %, mean±SD				
Extensive Metabolizer	76±23.6	73.6±21.8	85.6±10.2	85.6±9.9
Intermediate Metabolizer	72.4±26	72.4±26.2	76.9±10.4	75.4±10
Poor Metabolizer	65.6±29.9	65.9±29.6	71.7±11.2	63.2±12.1

The mean TTR for the PG arm was significantly higher than the STD arm in
the CoumaGen-I Simulation 2 (78.8% versus 73.7%; $P< 0.05$, respectively),
demonstrating that the Wilson protocol, which adjusts dose based on
percentage change, predicted better management of the clinical avatars and
was able to achieve a stable TTR for a longer period of time.

The CoumaGen-I Simulation 2 PG protocol resulted in a higher mean TTR
across all the genotype subsets than the corresponding CoumaGen-I
Simulation 1 PG protocol (Table 5.2). For all patients, the difference between
the STD arms in the CoumaGen-I Simulation 1 and 2 was 3.1%, indicating
similar TTR results despite different protocols. Conversely, the difference in
TTR for the PG arm was 6.8% higher, indicating that the Wilson protocol was
more accurate at maintaining a therapeutic dose within the 90-day clinical trial
time window. The CoumaGen-I Simulation 2 also exhibited a smaller TTR
standard deviation for each genotype subset than the CoumaGen-I Simulation
1, indicating that the INR range was better controlled by the use of the Wilson
protocol.

5.2.4.) Conclusion

This study showed that the clinical trial simulator was useful to study and

evaluate anticoagulation therapy options and "provide evidence to optimize the clinical trial for patient efficacy and reduced risk" (Fusaro, 2013).

5.3.) CoumaGen-II Clinical Trial Simulation

5.3.1.) Background and Objective

In this study, the accuracy of the clinical trial simulator of our *in silico* PC-CER framework was re-examined again by reproducing the CoumaGen-II clinical trial outcomes ("CoumaGen-II Simulation 1"), and then the simulator was used to evaluate a new dosing protocol ("CoumaGen-II Simulation 2"), to determine whether this new study design was significantly more beneficial for the same population.

5.3.2.) Methods

5.3.2.1) Study Design

- The original study. The original CoumaGen-II clinical trial (Anderson, 2012) comprised 2 prospective clinical trial comparisons: (1) a blinded, randomized comparison of 2 refined PG warfarin dosing algorithms and (2) a clinical effectiveness comparison of PG-guided therapy with use of either PG algorithm with a parallel, standard (STD) dosing (Figure 5.3). The primary end points of interest were Percent Time INR in Therapeutic Range (TTR) and also percentage of out-of-range (OOR) INRs during up to 90 day treatment period. In the original study, the inclusion criteria to recruit subjects were as follows: age ≥18 years, and an indication for initiation of warfarin anticoagulation. And the subjects with the following characteristics were excluded: women who were pregnant, lactating, or of child-bearing potential, those taking rifampin within 3 weeks, or patients with comorbidities precluding

standard dosing (e.g., creatinine >2.5, hepatic insufficiency, active

malignancy, advanced physiological age, expected survival <6 months);

noncompliance risk; and those deemed inappropriate for PG-guided dosing

for any other reason.

Based on power calculations, the minimum recruitment target for the

randomized, PG-guided comparison was set at 500 patients. All qualifying

parallel control patients were included, anticipated to number >=1000.

Eventually, 504 qualified subjects were recruited and underwent blind
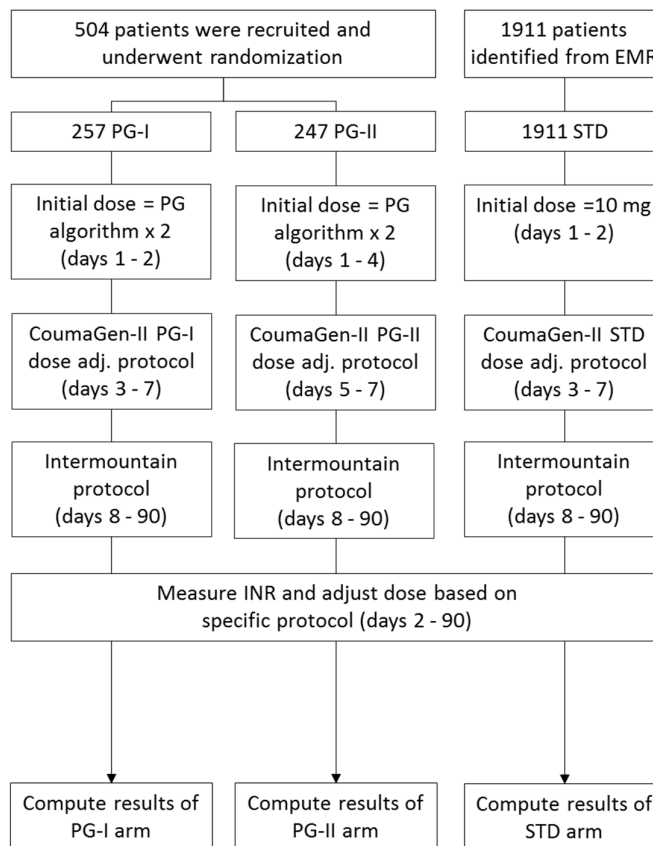
Figure 5.3. Study design for the original CoumaGen-II clinical trial.
PG: Pharmacogenetic arm, STD: Standard arm.

randomization to two PG arms; PG-1 and PG-2. The subjects for the

standard-dosing arm were retrospectively identified by a query to the EMRs of

3 hospitals. Patients ≥18 years of age initiating warfarin therapy with a

baseline and at least 1 follow-up INR level between days 3 and 14 were selected. The query resulted in identification and extraction of EMR records of 1911 subjects for the STD arm. Figure 5.3 illustrates the study design of the original clinical trial.

- The simulation study. In this study, the two following simulations were designed and conducted: "CoumaGen-II Simulation 1" and "CoumaGen-II Simulation 2". For each of these, with the use of a BNM method described in section 5.2.1.1, a sufficient number of clinical avatars was created to conduct 1000 simulations for each of the three arms; PG-1 (n=257000), PG-2 (n=247000), and STD (n=1900000). For each of the 1000 simulations, 257 avatars for the PG-1 arm, 247 avatars for the PG-2 arm and 1099 avatars for the STD arm were randomly recruited from their associated generated clinical avatar populations. Then, following the specific dosing protocol for each arm explained in the following sections, daily dose and INR for each avatar for that arm for 90 days were predicted.

CoumaGen-II Simulation 1 followed the dosing protocol as specified in the original CoumaGen-II clinical trial (Anderson, 2012; Figure 5.4). For the STD arm, it was assumed that retrospectively selected subjects in the original CoumaGen-II trial received warfarin standard initial dose of 5mg/day for days 1 and 2 and then the doses were adjusted using the same standard INR-based dose-modification algorithm developed and promoted by Intermountain. The CoumaGen-II PG-1 and PG-2 arms used modified versions of the IWPC (2009) dosing algorithm which required both clinical and genotype (i.e., CYP2C9 and VKORC1) information to calculate doses for days 1 and 2. Like the original study, the initial doses were followed by (a) a dose adjustment

algorithm based on INR according to the Kovacs (2003) protocol for days 3 to

7 for PG-1 arm and (b) a dose adjustment algorithm based on Lenzini (2010)

PG-based dosing algorithm for days 3 through 7 for PG-2 arm. For days 8 to

90 in both PG-1 and PG-2 arms, the Intermountain Healthcare warfarin dosing

protocol were used.
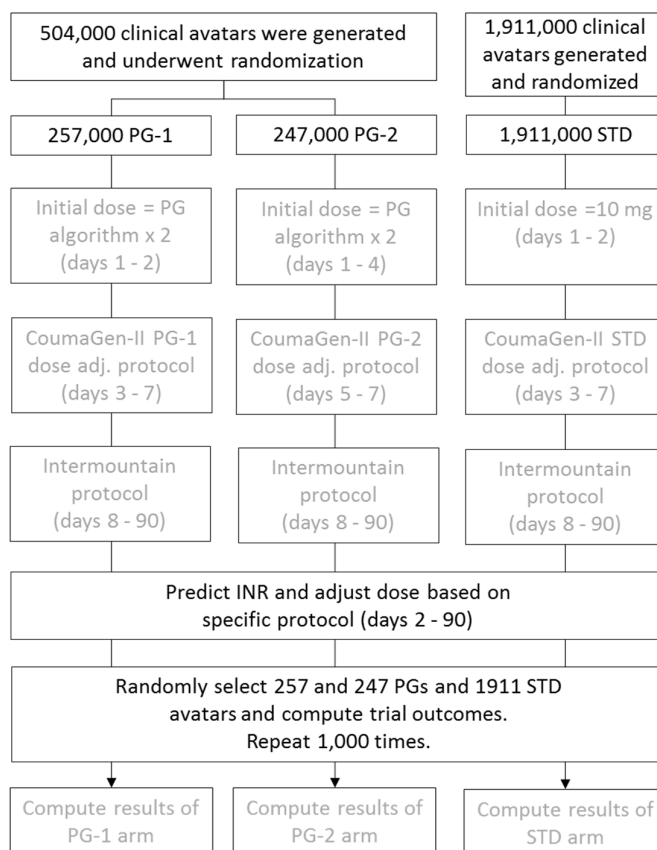
Figure 5.4. The study design of the CoumaGen-II Simulation 1. Black text represents new features compared to the original CoumaGen-II study, whereas gray text represents those features in common with the original study as depicted in Figure 5.3. PG: Pharmacogenetic arm, STD: Standard arm.

The design of the CoumaGen-II Simulation 2 was different from the original

clinical trial. It was conducted using a third dosing protocol (nomogram)

offered by Wilson (2007). In this study, the same initial doses for days 1 to 2

and the same adjustment doses for days 3 through 7 were used as in the

original CoumaGen-II trial. For days 3 to 90, the Wilson protocol was used to

increase or decrease the dose proportionally based on low or high INR values, respectively. The following figure (Figure 5.5) illustrates the design of the CoumaGen-II Simulation 2.
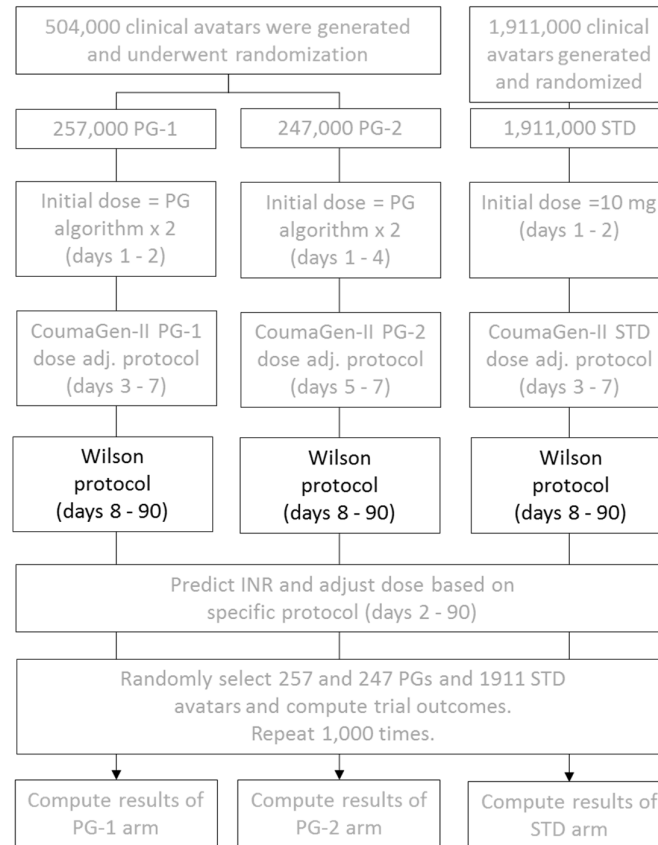
Figure 5.5. The study design of the CoumaGen-II Simulation 2. Black text represents new features compared to the original CoumaGen-II study, whereas gray text represents those features in common with the original study as depicted in Figure 5.3. PG: Pharmacogenetic arm, STD: Standard arm.

5.3.2.2) Clinical Avatar Populations

Similar to CoumaGen-I clinical trial simulation study (Section 5.2), for the BNM used for creating clinical avatars in this study, the same considerations were made as the original CoumaGen-II's individual patients' information were not available. Taking into consideration the subjects' characteristics required for the dosing algorithms used in this study, the prior probabilities used for the study's BNM were based on the statistical characterization of the patient

population including age, sex, weight, height, race, body surface area (calculated from height and weight), smoking status, deep vein thrombosis status, amiodarone use status, and genotypes for CYP2C9*2, CYP2C*3, and VKORC1. The CoumaGen-II trial data used age, sex, weight, height, deep vein thrombosis, smoking status, and genotypes as their variables.

In the simulation study, some other required prior probabilities (e.g., amiodarone status and genotype distributions) for the BNM were estimated by using some sources such as data from the Centers for Disease Control and Prevention, the 2000 US Census or some evidence on genotype distributions used for STD arm clinical avatar populations (e.g., Scott, 2010). The other details which were taken into for development of BNM in CoumaGen-I simulation study were also applied to this study too. The BNM was implemented in TETRAD IV (Scheines, 1998) to produce the clinical avatar populations for the both Simulations 1 and 2.

5.3.2.3.) Measuring Outcome Metrics

The primary outcome metric was Percent Time INR in Therapeutic Range ("TTR") and Percent Out-Of-Range INR ("%OOR"). Consistent with the original CoumaGen-|I study, TTR was defined as the percentage of time an individual avatar had an INR between 1.8 and 3.2 during the 90-day simulation. Although our simulator could calculate daily INR values, only INRs on those days in which the INR would have been checked in the clinic according to the specific protocol were considered.

5.3.2.4.) Clinical Trial Simulations

As stated in previous sections, in the original CoumaGen-II randomized controlled trial, the 504 recruited subjects were randomly assigned to two PG

arms; PG-1 (n=257) and PG-2 (n=247). In addition, in the original study, 1911 subjects were retrospectively recruited from the EMRs for the STD arm. Accordingly, in the Simulation 1 and 2 studies, the three clinical avatar populations were randomly sampled to recruit 257, 247 and 1911 avatars for the PG-1, the PG-2 and the STD simulation arms. The number of created avatars provided the opportunity to create 1000 parallel arms for simulations. Then daily dose and INR for each avatar were predicted following a specific protocol for that arm for 90 days. The predicted doses, INR, INR-monitoring frequency, and population statistics for each clinical trial simulation and across all 1000 simulations were recorded to calculate and produce the mean, standard deviation, and probability value (unpaired t test) for TTR for each study arm and for the aggregated populations from the 1000 clinical trial simulations. The simulations produced predictions were then compared with the CoumaGen-II results. The simulations were implemented in R (R Core Team, 2013) and run on the University of Wisconsin-Milwaukee's high performance computing environment.

5.3.3.) Results

Table 5.3 demonstrates the characteristics of the clinical avatar populations generated for PG and STD simulation arms along with the characteristics of the PG and STD arms' populations in the original CoumaGen-II study.
A few tests were done to show that the clinical avatar populations generated based on the BNM were statistically similar to the original CoumaGen-II population. As depicted in Table 5.3, the distribution of the characteristics across the PG arms and STD arm are very comparable and statistical test show no significant difference (P >0.05). The variable dependencies

embedded in the BNM of the populations were also tested to see if they also persisted in the simulated clinical avatar populations. To do so, a log-linear model was fitted to all relevant associations and probability values by using the Pearson χ^2 statistic were calculated. The test indicated that there were no significant differences (P >0.05) between the dependencies in the clinical avatar populations and the original CoumaGen-II populations (Table 5.3). In addition, it was tested to make sure that the non-embedded dependencies in the BNM did not exist in the clinical avatar populations.

Table 5.3. Characteristics of the subjects in the original CoumaGen-II study versus that of Clinical Avatars generated for PG and STD arms. DVT: deep vein thrombosis; PG: pharmacogenetic arm, SD: standard deviation, STD: standard clinical arm.

Characteristic	CoumaGen-II Original			Clinical Avatars		
	PG-1 Arm	PG-2 Arm	STD Arm	PG-1 Arm	PG-2 Arm	STD Arm
n	257	247	1911	257	247	1911
Age, y (mean)	61.3	59.9	51.5	60.9	59.4	51.1
Male, %	46.3	47.6	48.5	46.9	48.3	49.1
Weight, kg, mean±SD	93.1±23.9	92.3±24.5	91.3±28	92.1±23.1	92.9±23.5	90.9±27.5
DVT, %	29.6	29.6	33.7	28.9	29.2	33.9
White, %	95.3	95.6	91.4	95.6	96.1	91.1
CYP2C9*2, %						
CC	79	77.6	NA	79.5	77.2	78.3
CT	17.9	20.3	NA	17.7	20.9	19.1
TT	2.7	1.2	NA	2.8	1.9	2.6
CYP2C9*3, %						
AA	86.3	86.1	NA	85.9	86.6	86.2
AC	13.3	13.1	NA	13.7	13.1	13.2
CC	0.4	0.8	NA	0.4	0.3	0.6
VKORC1, %						
G/G	32	38.1	NA	31.5	38.8	35
G/A	49.6	47.5	NA	50.3	46.7	48.6
A/A	18.4	14.3	NA	18.2	14.5	16.4

The accuracy of the clinical trial simulator of our *in silico* PC-CER framework was validated again by simulating the original CoumaGen-II clinical trial through conducting CoumaGen-II Simulation 1 (Figure 5.3 and 5.4). Similar to

the results of the original CoumaGen-II, the CoumaGen-II Simulation 1 study showed that the PG-2 dosing algorithm was noninferior compared with the PG-1 dosing algorithm at 1 and 3 months (P >0.05, Table 5.4). It also verified the same for %OOR INR at 1 and 3 months (P >0.05).

Through the second simulation, CoumaGen-II Simulation 2, it was tested if a modification to PG and STD arms of the original CoumaGen-II dosing algorithms would result in a significant change in outcomes of the three arms. As depicted in Figure 5.5, all model and simulation components of the CoumaGen-II Simulation 1 (i.e., clinical avatars, initial dosing, PK/PD parameters, and TTR outcome calculations) remained the same with the exception of the replacement of part of the original CoumaGen-II dosing protocol with the Wilson protocol (Wilson, 2007) for days 8 to 90. The simulation, CoumaGen-II Simulation 2, was run and the outcome metrics were calculated as in the Simulation 1 (Table 5.4).

Table 5.4. Time INR in Therapeutic Range (%TTR) and Out-Of-Range INR (%OOR) across the 1000 trials of CoumaGen-II Simulations 1 and 2.

Outcome Metric	CoumaGen-II Simulation 1			CoumaGen-II Simulation 2		
	PG-1 Arm	PG-2 Arm	STD Arm	PG-1 Arm	PG-2 Arm	STD Arm
TTR, %, mean±SD 1 month	64.6± 19.8	62.9± 17.6	41.5± 13.5	68.2± 10.9	67.4± 11.2	55.1± 13.4
TTR, %, mean±SD 3 months	66.1± 18.5	64.5± 19	42.3± 13.8	70.4± 9.3	69.4± 10.1	56.5± 11.8
OOR, %, mean±SD 1 month	35.4± 11.5	37.1± 11.4	58.5± 12.2	31.8± 8.6	32.6± 9.1	44.9± 7.8
OOR, %, mean±SD 3 months	33.9± 10.3	35.5± 10.7	57.7± 11.1	29.6± 6.4	30.6± 7.9	55.1± 7.8

The 1-month and 3-month mean TTRs for the PG-1 and PG-2 arms were significantly higher than the STD arm in the CoumaGen-I Simulation 2 (70.4% and 69.4% versus 73.7%; P< 0.05, respectively), demonstrating that the

Wilson protocol, which adjusts dose based on percentage change, predicted better management of the clinical avatars and was able to achieve a stable TTR for a longer period of time.

The CoumaGen-II Simulation 2 PG protocols resulted in a higher mean TTR across all the genotype subsets than the corresponding CoumaGen-II Simulation 1 PG protocol. The CoumaGen-II Simulation 2 also exhibited a smaller TTR standard deviation across all three arms than the CoumaGen-II Simulation 1, indicating that the INR range was better controlled by the use of the Wilson protocol.

5.3.4.) Conclusion

We have developed a pharmacogenetic clinical trial simulation framework for warfarin dosing and validated the framework against the CoumaGen-II clinical trial. We also demonstrated the utility of our framework by simulating the same clinical trials with the use of a relatively more aggressive dosing protocol and predict that the PG arms are likely to perform significantly better than the STD arm. The framework provides an opportunity to assess alternative strategies such as different dosing protocols, study population, or outcome metrics before applying them in real world.

5.4.) Aurora Health Care Anticoagulation Therapy Simulation

5.4.1.) Background and Objective

The complexity of anticoagulation therapy and various existing treatment options especially the new PG-based dosing protocols create a serious barrier to hospitals of identifying and adopting an "optimal" anticoagulation treatment

plan for their heterogeneous patient population. To address this challenge, we designed a study to take advantage of our pharmacogenetic PC-CER *in silico* framework to simulate a number of anticoagulation therapy scenarios for the state-wide AHC patient population. Our approach included extraction of representative warfarin patient data from WiAD, collect and codify AHC's warfarin dosing protocols, run clinical simulations to compare the predicted outcomes of different dosing protocols including the AHC's, and then identify the overall best population-wide treatment plan for different subpopulations.

- AHC warfarin Dosing Protocols

AHC has an institution-wide standardized warfarin best practice treatment protocol that has been in effect for the last decade in different treatment facilities of the institute. The AHC protocol called "Aurora Anticoagulation Clinic Guideline for Ambulatory Warfarin Management" (AACG–AWM) has some special characteristics compared to the other dosing protocols used in our previous simulation studies. In the following, some details about the protocol are offered.

AACG-AWM has two dosing protocols where each has three main components: initial dose (days 1 and 2), 1st INR in range, and INR-based Dose Adjustment. According to the protocol, a fixed dose of warfarin is prescribed for days 1 and 2 following the instructions in table 5.5.

Table 5.5. AHC warfarin dosing protocol - AACG-AWM.
Initial dosing for days 1 and 2.

Patients sufficiently healthy to be treated as outpatients	= 10mg/day
Patients 65+ years of age	≤ 5mg/day
Patients at any age and with multisystem disease, known liver disease, taking drugs that are likely to increase warfarin effect, have had prior at-goal treatment response with low doses or have baseline INR readings above 1.1	≤ 5mg/day

Two days post warfarin initiation, INRs are monitored daily or every other day until the INR >= 2.0 or as indicated by referring physician. When this is achieved INR testing follows the chronology below (Table 5.6).

Table 5.6. AHC warfarin dosing protocol - AACG-AWM. 1st INR in range.

Number of days the INR and warfarin dose remain stable and therapeutic	Days until the proceeding INR test
2	3-5
7	7
14-21	14
28-35	28
42	42
56	56
84	84 (absolute max number of days between tests)

Based on this protocol, any adjustment of dosage is based on INR records in last few days prior to dose adjustment. This aspect makes the AACG-AWM different from other doing protocols such as CoumaGen-I and -II (Anderson 2010a, 2012b) as they take into consideration only the INR value that was recorded just prior to dose adjustment. Figure 5.6 provides details on the third component of the AACG-AWM protocol which instructions on INR-based dose adjustment.

The AACG-AWM protocol has some other components such as the following which were not taken into consideration in the process of codifying the protocol as they either were not providing objective measures or the WiAD dataset did not provide relevant information on them.

- Drug Interaction Considerations with Warfarin

- Herbal/Natural Medicine Interaction Considerations with Warfarin

- Other Factors That May Affect INR.
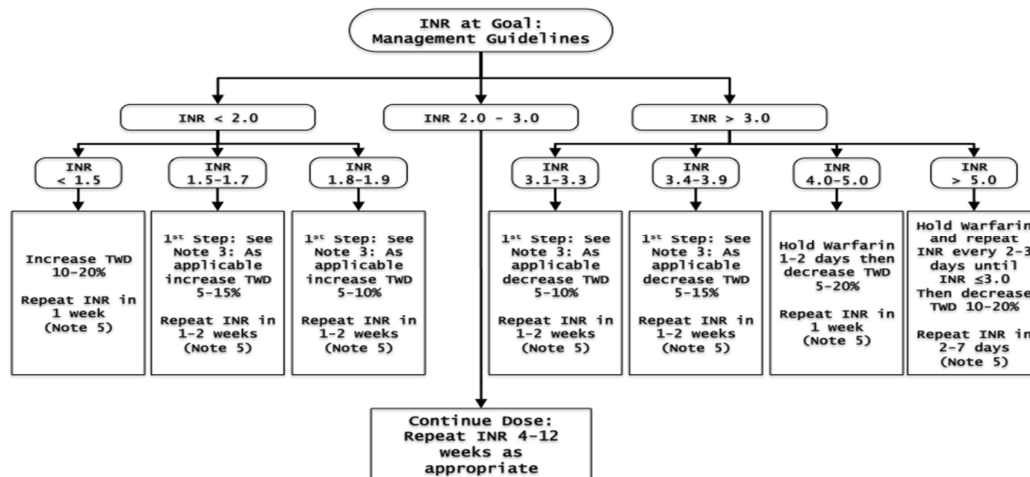
Figure 5.6. AHC warfarin dosing protocol - AACG-AWM. INR-based dose adjustment. Adjusting dosage for patients with INR goal of 2.5 (range 2.0 – 3.0).

5.4.2.) Methods

In this study, our primary hypotheses were as follows:

(1) Any warfarin dosing algorithm that provides some measure of personalization will demonstrate improved predicted clinical outcomes (as defined by TTRs and %OOR INRs within 90 day period) compared with current AHC 'best-practice' dosing algorithm across the entire AHC population. We defined personalization as any demographic, clinical or genetic variable applied across a medication dosing protocol.

(2) Predicted clinical outcomes (as defined by TTRs %OOR INR's within 90 day period) will improve proportionally to the degree dose personalization across the entire AHC population.

5.4.2.1.) Study Design

To test these hypotheses, we designed a five-arm simulation study (Figure 5.7) in which the study arms (AHC, Clinical, PG-1, PG-2, and PG-3 arms)

have different levels of personalization depending on required subjects' characteristics for warfarin dose adjustments. As depicted in Figure 5.7, each arm's dosing protocol composed of three components. These components are fixed, clinical-based or PG-based. In the first arm (AHC arm), the initial dosing was solely based on the age and health condition of the subjects whereas in the other arms it was based on either clinical characteristics (i.e., Clinical arm) or clinical characteristics along with genotypic profile of the subjects (i.e., PG-1, PG-2 and PG-3 arms). In the first three arms, the dose adjustment algorithm used was the AHC one. For the arms PG-2 and PG-3, the dose adjustment algorithm was a PG-based one used in EU_PACT clinical trial (Pirmohamed, 2013). Although the maintenance dosing algorithms (days 6 or 8 through 90) for all of the five arms were non PG-based, the AHC maintenance dosing algorithm was used for the first 4 arms and the last arm's maintenance dosing was based on the Intermountain dosing algorithm (Anderson, 2012).
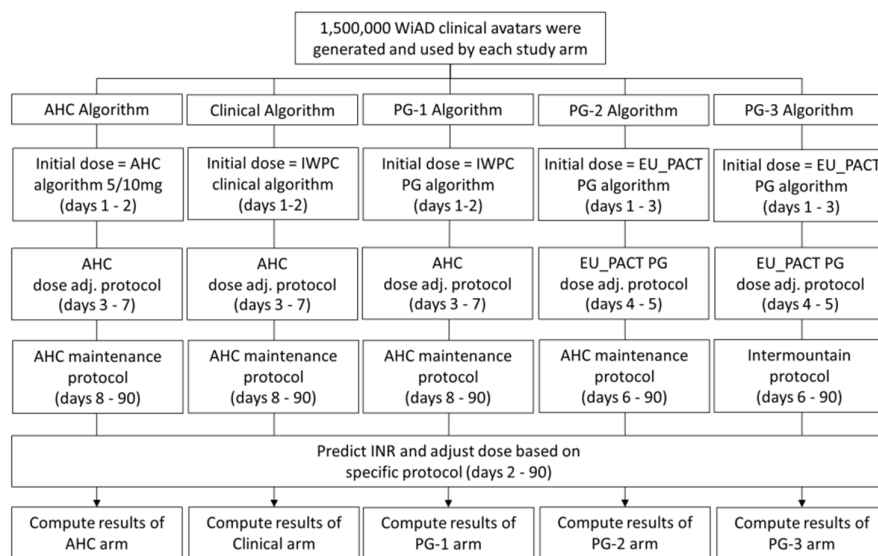

Figure 5.7. Study design for the five simulation trial arms.

- Study Population

Using WiAD-Miner, we identified and extracted the EMR records of all the subjects who were under warfarin at AHC and had complete demographic and clinical records. This query resulted in 14,206 subjects.

- Study Power Analysis:

Taking advantage of our *in silico* framework, we had the opportunity to test our hypotheses against a wide variety of subpopulation such as minority subpopulations which had not participated in significant numbers in recent or past warfarin clinical trials (e.g., African American). With these subpopulations identified, we calculated the statistical characteristics of each subpopulation. Accordingly, we found out that the smallest subpopulation of interest was African-American male (i.e., 1.9% of whole extracted WiAD warfarin population). A power analysis was conducted to calculate the sufficient number of avatars to reject our null hypothesis for this subpopulation (African-American males). Our power analysis showed that we needed to have at least 250 African American male avatars within each arm (Significance level 0.05 two-sided, SD 0.2, Power 0.9, INR Improvement [Difference in means] 0.06). Consequently, the calculation resulted in a minimum number of 13,015 avatars for each arm.

5.4.2.2) Clinical Avatar Populations

For the purpose of this study, we applied the clinical avatar modeling method described in section 3.1.1 to the WiAD warfarin population data to generate clinical avatars. As mentioned in chapter 3, our BNM pipeline consists of four broad sections: (1) Data preprocessing, and Knowledge aggregation; (2) Develop Directed Acyclic Graphs (DAGs) via Ensemble Learning; (3)

Estimating conditional probabilities via ensemble learning; (4) Validation of the BNM and imposition of variables. In the following section, the application of our methodology to the WiAD warfarin population is described in detail. Following our clinical avatar model described in chapter 3, clinical avatar modeling and generation were done for AHC warfarin patient population (i.e., WiAD warfarin cohort). As mentioned in chapter 3, our BNM pipeline consists of four broad sections: (1) data and knowledge aggregation and preprocessing; (2) semi-supervised Bayesian pattern search to develop a Directed Acyclic Graph (DAG); (3) applying the DAG to estimate joint conditional probabilities and deriving an instantiated model; (4) the BNM must be validated against a subset of data not used in developing the BNM. In the following section, our methodology for the WiAD warfarin cohort is described in detail.

Section 1: Data preprocessing, and Knowledge aggregation

This section has two branches performed in parallel. In Part I we accumulated domain knowledge (i.e., expert knowledge or literature-defined knowledge) to better understand both the specifics of our data such as the way the data was gathered, the semantics of the data dictionary and any measurement error as well as the general relationship between the variables as found in literature review. In Part II the patient data from WiAD was characterized and prepared (i.e., data preprocessing). The data preparation is a multistep process as described in Figure 5.8. Once the data has gone through the data preprocessing procedure it is aggregated into a data wrapper.
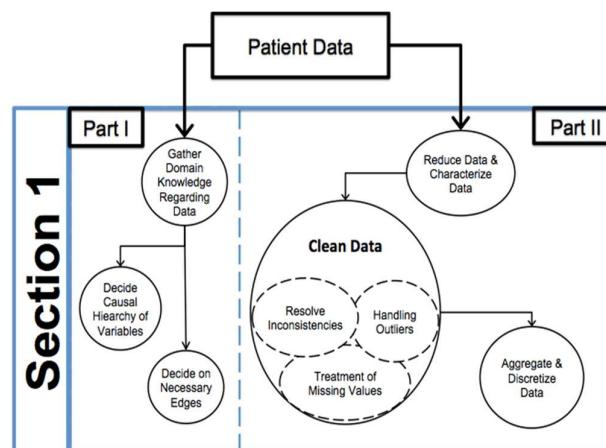
Figure 5.8. Step-by-Step procedure for Section 1

Part I:

We undertook a complete literature review searching Google scholar and
PubMed using key words such as "Race and Smoking" and "BMI and Age",
similar to those variables found in the Aurora Data. We filtered for articles that
provide explicit results on populations determined to be similar to the AHC
patient population. We then filtered for results that linked the same specific
variables as those found in the WiAD data through an asymmetric conditional
probabilistic relationship. In Figure 5.9A and 5.9B we demonstrate the results
of the literature review and then how these relationships are implemented in
the TETRAD program (Scheines, 1998). Figure 5.9A shows a total of 7
asymmetric conditional probabilistic relationships discovered through literature
review. An asymmetric conditional probabilistic relationship is not strictly
causality nor do we require this strong assumption to be true in the
development of clinical avatars. Each of the required edges, shown in the
green directed arrows are each related to evidence cultivated through the
literature review that were determined to relate direct to the patients in the
WiAD database (Fiore, 1989; Kuskowska; 1992; Marot, 2011).

In Figure 5.9B we see the description of knowledge tiers within TETRAD. The tiers allow use to inform the algorithm that certain relationships are both forbidden and/or asymmetric. As implemented for the AHC data, we suggest that IF a conditional dependency is discovered between "HEIGHT" and
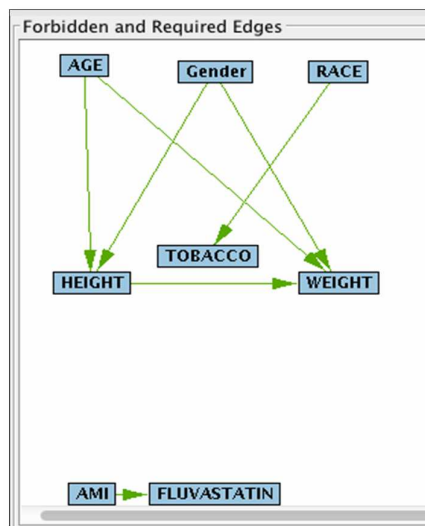
Figure 5.9A. Domain Knowledge expressed as required causal relationships in green arrows
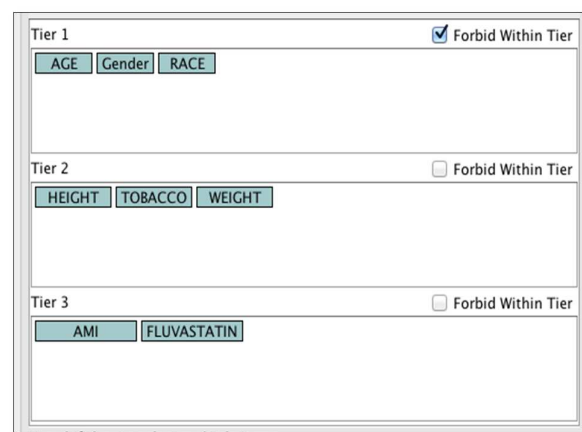
Figure 5.9B. Domain Knowledge expressed as causal tiers of hierarchy

"RACE", Then the relationship must be such that "RACE" given its position in "Tier 1" is the causal parent of "HEIGHT" given its position in "Tier 2". Similarly any potential relationship between a higher tier variable must be directed towards a lower tier variable and never in the reverse direction.

Part II:

Data Discretization and Aggregation:

Often healthcare data includes a mix of categorical and discrete data, such as race and gender, and continuous data such as age and height. The Bayesian search algorithms we employ require training data that is either entirely continuous and normally distributed or entirely discrete. The WiAD data

included a combination of both continuous and discrete variables, therefore the continuous variables were discretized. The process of discretizing continuous data is a ubiquitous data preprocessing technique that must balance information loss inherent in the process with the benefits of greater processing efficiency. There are numerous discretization methods and the choice can impact both the posterior probability estimation as well as the discovery of inherent causal structure in the underlying graph. We employed a common unsupervised method, EqualWidth that has demonstrated its ability to produce accurate data mining results for Bayesian search algorithms when compared to other techniques (García-Laencina, 2013).

Once the data has been fully preprocessed it is loaded into TETRAD the program we use to implement the BNM search and relationship discovery. The final step before completing the Bayesian search, is to load the processed data as a data wrapper. A data wrapper converts the data from a flat structural file to a relational file that describes the data as tuples. Data wrappers are standard in data mining practices because it allows for maximum dimensionality to be passed through the search algorithm.

Section 2: Directed Acyclic Graphs via Ensemble Learning

Bayesian search algorithms are based on a number of assumptions that often times breakdown when applied to real data. We address the weakness of the Bayesian search algorithms, along with certain bias and variance problems within the aurora data by eliciting established methods for using ensemble learning such as described in Part III section 2. The step-by-step logic of section 2 is described in Figure 5.10. We apply an ensemble learning technique called, bootstrap aggregation (Bagging) to address the instability of

Bayesian classifier searches. From one set of training data, the data is resampled at random.
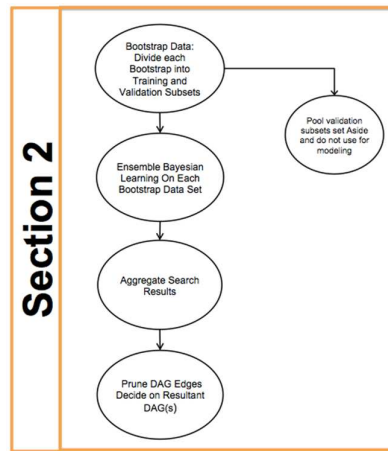
Figure 5.10. Developing a DAG from Training Data

The choice of bootstrap sampling technique (e.g., Parametric vs. classic bootstrap) can ultimately impact the quality of the model. We implemented a classic (Unweighted) bootstrap because imputation and missing values were non-existent, and there was limited bias within the selected EMR dataset. The classic bootstrap is completed from the original data wrapper after it has been loaded into TETRAD. In Figure 5.11, the data is labeled as "WiAD_Warfarin_Cohort Data" before entering the pipeline. This data wrapper is then aggregated into five distinct bootstraps and subdivided into 80% and 20% subsets with 80% used as the training data. The bootstraps are labeled as "Training_DATA1… Training_DATA5" in Figure 5.11. The bootstraps were resampled to the same size n=14,206 as the original dataset. The 80% training data, n=11,365, was used as input into 5 Bayesian searches, labeled "Search1…Search5". Knowledge that was previously loaded during the preprocessing phase is also applied evenly across the search algorithms to constrain the search space.
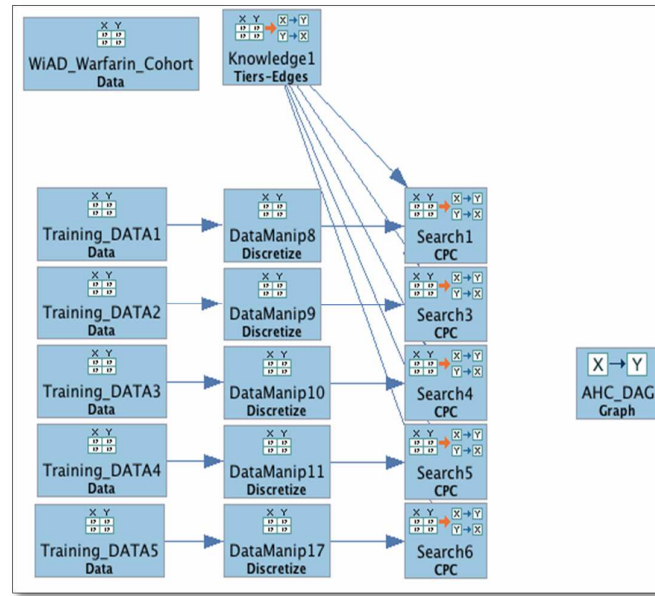
Figure 5.11. Bootstrapping and performing
searches constrain by knowledge in TETRAD

Additionally, in experimentation with EMR data we employ a "repeated leave

one out bootstrap aggregation" method (Clyde, 2004; Jiang, 2007). Holding

out data for cross validation is considered a data mining 'best practice'

(Belazzi, 2004). However, when the sample size is small and outliers are

important considerations in the data, holding out any amount of data reduces

variance in the model. Therefore, it is recommended to perform the bootstrap

sampling of the original data prior to the dividing the data between training

and validation subsets.

Each resampled training data subset was used as input with the previously

described domain knowledge to a Bayesian search algorithm. The results of

each search are aggregated and the results are pruned to yield a resultant

DAG. We employed the Bayesian search algorithm called the Conservative

PC algorithm (CPC) found within the TETRAD publicly available software. The

CPC algorithm is a variant of the PC algorithm that has an additional step that

provides additional autonomous arrow directionality. The PC and CPC

algorithm perform conditional independence tests between each variable except those required or forbidden within the knowledge box. The conditional independence tests are then applied across the data wrapper to discover potential causative relationships. Additional details, the pseudo code and inherent assumptions of the PC and CPC algorithm can be found in (Sprites, 2000) and within TETRAD. All searches were completed with an α=0.05. We generally recommend using multiple Bayes search algorithms for each bootstrap of data to address the inherent instability in the search results. However, in application to the WiAD dataset, we determined that the CPC algorithm sufficient for the following reasons: There was limited number of dimensions (i.e. variables) in the data in proportion to the sample size. Also, because all missing values were deleted from the training dataset, a major source of algorithm instability was addressed in preprocessing. Lastly, CPC was determined to best fit the data, and there was limited variation between bootstrapped search results.

After the 5 CPC pattern searches were performed on the data, the DAGs were converted into matrix with the originating state comprising the columns and the destination states the rows. Each directed arrows, or edges of the DAGs were cataloged in a matrix for comparison. Edges are selected based on those that received 50% or more commonality and/or are supported by evidence discovered prior to performing the search. The results of the Five CPC searches are shown in Figure 5.12. The weight of the arrow corresponds to the number of times the search output confirmed an existing edge. The two arrows that are dashed from Gender → Tobacco and Race → Weight had fewer than 50% search results and therefor were not included in the

aggregated DAG. The resultant DAG shown in Figure 5.12 is then used in parallel for section three, labeled "AHC_DAG" in Figure 5.14.
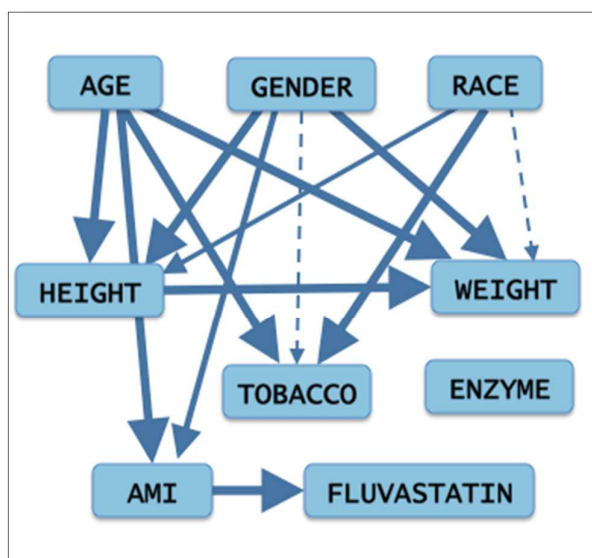
Figure 5.12. Results from 5 CPC searches.

Section 3: Estimating Conditional Probabilities via Ensemble learning

The step-by-step logic of section 3 is described in Figure 5.13. The dashed
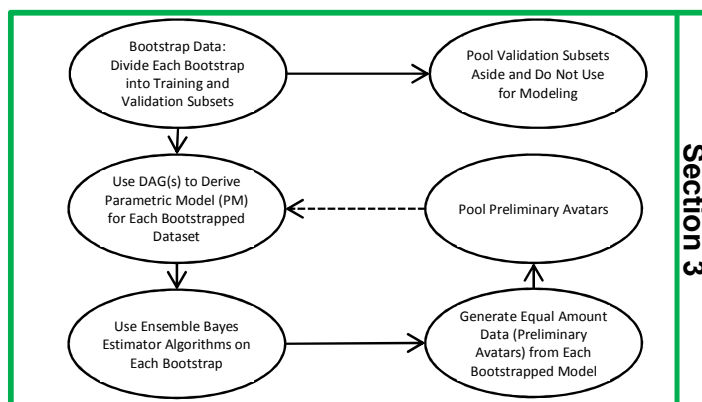
Figure 5.13. Developing a DAG from Training Data

arrow indicates a final recursive step described at the end of this section. The first step in section 3 is to again bootstrap the original data set the same number of times as performed in section 2. The bootstraps were resampled to the same size n=14,206 as the original dataset. Because the justification for selecting a bootstrap procedure is the same in this section as for section 2,

we performed 5 classic bootstraps with replacement holding out 20% of each bootstrap for validation in section 4, (identical bootstrap to section 2). The bootstraps were performed in TETRAD as seen in Figure 5.14 represented by ("Training_Data6… Training_Data10). The DAG derived from section 2, "AHC_DAG" was used to develop a parametric model (PM) for each of the bootstraps.
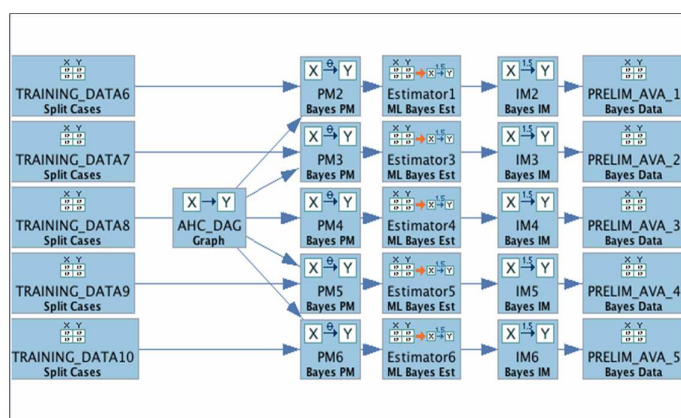
Figure 5.14. Parameter learning in TETRAD

The parameters, or conditional probabilities of the Bayes net were estimated for each bootstrap using the ML Bayes estimator for each bootstrap as shown in Figure 5.14 noted by the "Estimator1… Estimator5"). At this stage the algorithm produced a partial failure. An example of this problem is highlighted in Figure 5.15 for particular parameters of "WEIGHT" within the BNM. The partial failure of the estimator algorithm is highlighted by the red rectangles within the partial estimated Bayes net. In the resultant DAG from section 2, "WEIGHT" is a descendent of three Variables, "HIEGHT", "AGE" and "GENDER". The partial failure is associated with the problem of insufficient statistics to provide estimates within those particular parameters highlighted in red. In this case, the training data did not provide the estimation algorithm with

any cases from these particular parameters (e.g, there existed no patients with the characteristics "Gender=Female, Age=0, Hieght=0" within the
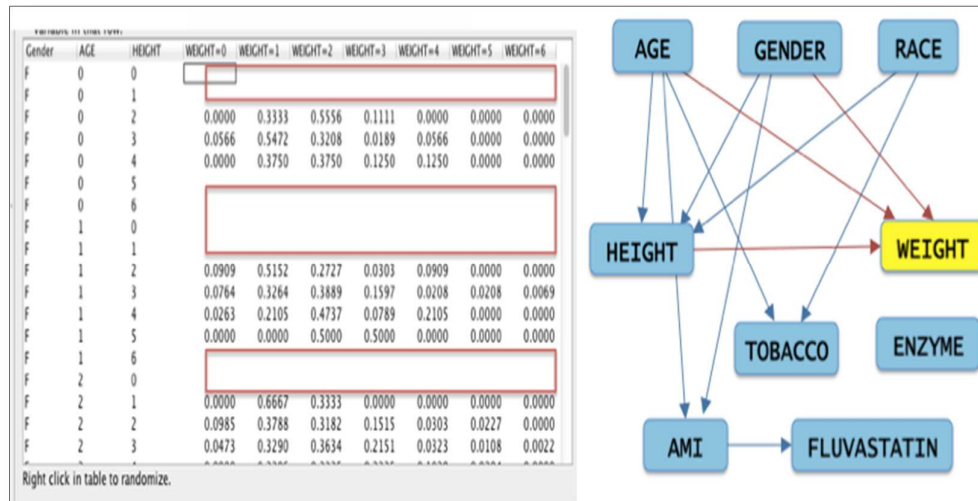
Figure 5.15. Partial failure of Estimator Algorithm for the variable "WEIGHT" bootstrap of the training data).

To address this situation, and satisfy the parameter learning for this BNM, we simplified the model by removing edges related to "WEIGHT" according to the four-step logic pattern found in the general description of the method. In this case, there were three edges related to "WEIGHT". Since all three edges were retuned every search result performed in section 2, we began by increasing α from a value of α=0.05, to a more stringent value of α=0.00001. We then used the domain knowledge aggregated in section 2 to delete further edges stepwise until sufficient cases existed within training data to calculate conditional probabilities. In doing so we increase the number of cases in the training data to calculate the conditional probability for all relevant states from the parent nodes. The resultant DAG, named "SIMPLIFIED_AHC_DAG" was used to derive PM and estimate the conditional probabilities of the simplified Bayes Net shown in Figure 5.16. Figure 5.17 highlights our example of the

variable "WEIGHT". We see that the edges "AGE" and "GENDER" were removed while the edge between "HEIGHT → WEIGHT" remained. The conditional probability for each discretized category of weight was then estimated strictly based on the discretized categories of height.
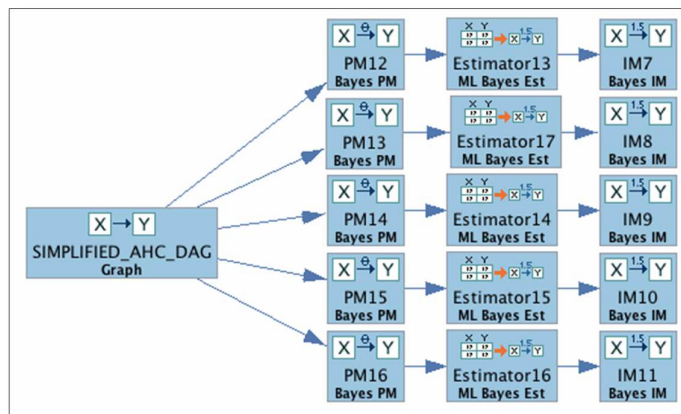
Figure 5.16. Parameter learning with insufficient statistics

The conditional probabilities derived from the simplified model were then used as input for the original parametric model. For example, within the more complicated Bayes net we experienced a partial failure when the training data did not have any counts for patients that have "Gender=Female, Age=0, Height=0". The Estimated conditional probability for this parameter was derived from the simplified Bayes net – using weight estimations strictly from its probabilistic relationship to Height. This procedure was repeated for "Tobacco" and "Height" until all parameters within the original Bayes net were satisfied with conditional probabilities for each bootstrap of data.

Once the parameters of each Bayes net were satisfied within an instantiated model (labeled IM within Figure 5.16), 100,000 preliminary avatars were generated from each trained model. The preliminary avatars were then pooled in equal proportion into a set of "POOLED_PRELIM_AVATARS". We elected

to use an equal weighting scheme because, similar to the reasons for performing a classic bootstrap, the bias was considered minimal. The pooled preliminary avatars were then passed through a final round of estimation. The resultant "AHC_DAG" developed in section 2 was then used to derive the PM that was used as input to the estimator. Because of outliers within the data this second round of estimation again faced the problem of insufficient statistics that resulted in partial failure of the estimator algorithm. This, despite the large sample size of pooled preliminary avatars passed through the Estimator algorithm. We addressed the problem in an identical manner to the first round of parameter learning. We used a simplified DAG to reduce the conditional parents and substituting the conditional probabilities for the simplified state for those states within the parametric model that suffered from the partial failure. We used the same simplified DAG previously developed and shown in Figure 5.17.

HEIGHT	WEIGHT=0	WEIGHT=1	WEIGHT=2	WEIGHT=3	WEIGHT=4	WEIGHT=5	WEIGHT=6
0	0.6667	0.3333	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.3247	0.5584	0.0909	0.0260	0.0000	0.0000	0.0000
2	0.2372	0.5187	0.1915	0.0437	0.0074	0.0016	0.0000
3	0.0712	0.4949	0.3224	0.0922	0.0150	0.0034	0.0010
4	0.0055	0.2779	0.4868	0.1700	0.0447	0.0136	0.0015
5	0.0015	0.1090	0.4582	0.2881	0.1075	0.0299	0.0060
6	0.0000	0.2500	0.2500	0.3750	0.1250	0.0000	0.0000

Figure 5.17. Simplified DAG and subsequent parameter learning

The parameters were then aggregated in the "TRAINED_AHC_BNM" shown in Figure 5.18. This BNM was then used to develop 100,000 clinical avatars via Monte Carlo simulation techniques that were used as input to section 4 for validation purposes.
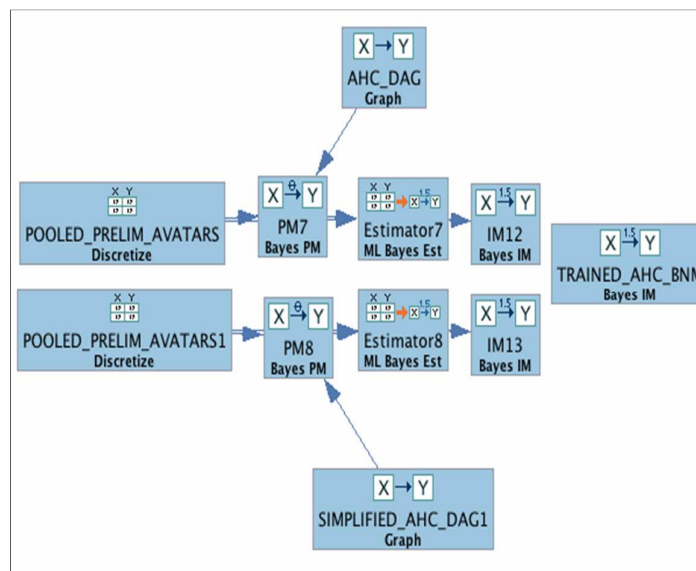
Figure 5.18. Final round of parameter learning

Section 4: Validation of the BNM and Imposition of Variables

The first step of this section involves validating the model with some set validation data shown in Figure 5.19. We performed a "repeated hold-out classic bootstrap" in both section 2 and section 3. A total of 5 classic (i.e., with replacement) unweighted bootstraps were performed in section 3. The bootstraps were resampled to the same size n=14,206 as the original dataset. Each described previously, each bootstrap was randomly divided in 80% and 20% subsets. Therefore a total of 5 validation data sets each with the same size, n=2,841 were aggregated into a data wrapper labeled "POOLED_VALIDATION_DATA" in Figure 5.21. This set of validation data

was then used to compare the 100,000 clinical avatars for validation

purposes.

There are two types of comparisons used to ensure that the simulated

population is representative of the original data. The first is univariate
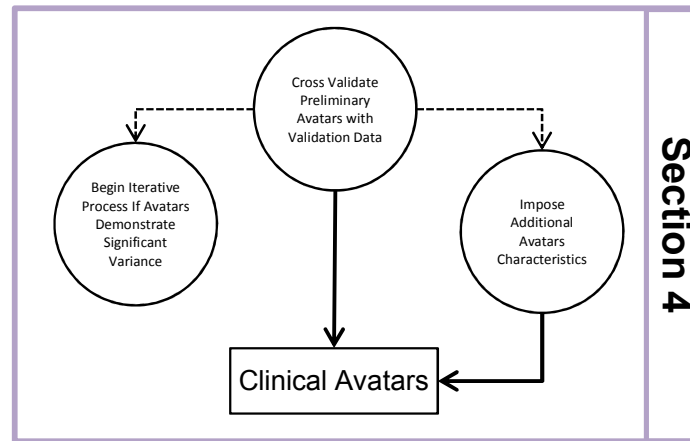
Figure 5.19. Validating the BNM and generating Clinical Avatars

distribution, or the comparative frequency of a single attribute between the

training data and clinical avatars. The second comparison is the bivariate

frequency distribution between the training and clinical avatar data sets.

Variables in which the Bayesian algorithm determined a causal connection

and are "d-connected" are plotted in frequency histograms. We found no

significant variance in either the univariate or bivariate distribution, the

process is refined until there is no significant difference. The comparative table

is provided in the results section.

The final step in the development of clinical avatars involves imposing any

additional characteristics on the BNM validated against the pooled validated

data. In Figure 5.20, we demonstrate how two genotypes, CYP2C9 and

VKORC1 are imposed on the BNM derived from the WiAD database.
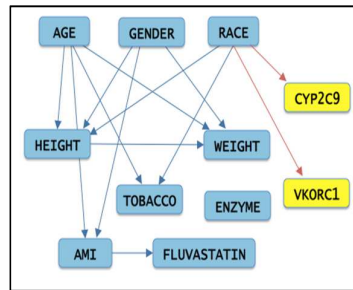
Figure 5.20. Final AHC DAG following Genotype Simulation.

The probabilistic dependent relationship between "RACE → CYP2C9" and

"RACE → VKORC1" are demonstrated in (Scott, 2010). Additionally, the

parameters that describe the conditional probability parameters are described

in (Scott, 2010). The validated structure and parameters that include all

clinical avatar parameters including genotype are demonstrated in the

Instantiated BNM titled "AHC_BNM_GENOTYPE" in Figure 5.21. This model

was then used to generate 1,500,000 clinical avatars via Monte Carlo

simulation techniques that were entered into the simulation platform.
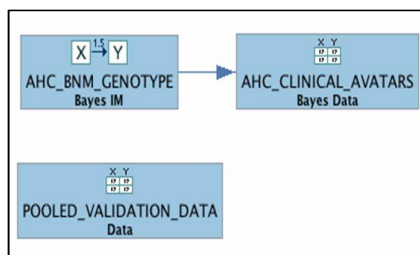
Figure 5.21. Section 4 as implemented in TETRAD

5.4.2.3.) Measuring Outcome Metrics

In this study, we calculate a number of outcome metrics. A primary outcome

metric is Time in Therapeutic Range (TTR) for patient INR. There are two

methods for calculating TTR, such as INR check points and linear

interpolation (Rosendaal, 1993). Additional outcome metrics include the following: 1&2.) Percent time INR higher than therapeutic range using both INR check points and linear interpolation methods, 3&4.) Percent time INR lower than therapeutic range using both INR check points and linear interpolation methods, 5.) Number of INR predictions, 6.) First day INR higher than therapeutic range, 7.) Percent INR in therapeutic range by day 5, 8.) Percent INR in therapeutic range by day 9, 10.) number of dose adjustment, 11.) relative risk of ischemic stroke, and lastly 12.) relative risk of intracranial hemorrhage.

5.4.2.4.) Clinical Trial Simulations

For each study arm, a 90 day simulation was performed for each Clinical Avatar using each of the five PG and non-PG protocols. All results were stored in a structured format representing the 100 x 90 day study simulations for each clinical avatar in the study subpopulations. Simulation records included: clinical avatar record, simulated INRs and dose values (1 per day for each of 90 days) and calculated outcome metrics. All simulations were implemented in R (R Core Team, 2014) and performed on the UWM's high performance computing research cluster Avi (UWM, 2014).

5.4.3.) Results

In the following table we present the statistical characteristics of the clinical avatars and the WiAD warfarin study population. The statistical analysis indicates no significant difference between these two populations by the characteristics.

Table 5.7. Characteristics of the WiAD warfarin study population versus WiAD warfarin clinical avatar population. SD: standard deviation.

Characteristic	WiAD Warfarin Population	WiAD Warfarin Clinical Avatar Population
Age, y, mean±SD	67.3 ± 14.43	67.2±14.47
Weight, lb, mean±SD	199.24 ± 54.71	199.24±54.6
Height, in, mean±SD	66.78 ± 4.31	66.53±4.32
Gender, %		
Female	53.14	53.10
Male	46.86	46.90
Race, %		
White	95.17	95.19
Black or African-American	4.222	4.202
Asian	0.3378	0.4010
Am. Indian/Alaskan	0.1759	0.1890
Pacific Islander	0.0007	0.0001
Tobacco, %		
No	90.33	90.67
Yes	9.66	9.33
Amiodarone, %		
No	88.45	88.49
Yes	11.54	11.51
Fluvastatin, %		
No	99.97	99.98
Yes	0.03	0.02
CYP2C9*2, %		
*1/*1	NA	67.39
*1/*2	NA	14.86
*1/*3	NA	9.25
*2/*3	NA	1.97
*2/*2	NA	6.51
*3/*3	NA	0
VKORC1, %		
G/G	NA	38.36
G/A	NA	44.18
A/A	NA	17.45

The following figure demonstrates the mean predicted TTR of the 100 simulations for each arm. The comparison of the results of linear interpolated TTR across the whole clinical avatar population for all 5 arms is shown in Figure 5.22. The PG-1 arm produced the highest mean predicted TTR, at

77.43%. The Clinical arm and PG-3 arm produced similar but significantly different (p<0.05) mean predicted TTRs at 69.76% and 67.4%, respectively. PG-2 and AHC arms produced significantly inferior mean predicted TTRs (P<0.05) at 62.99% and 57.16% respectively.
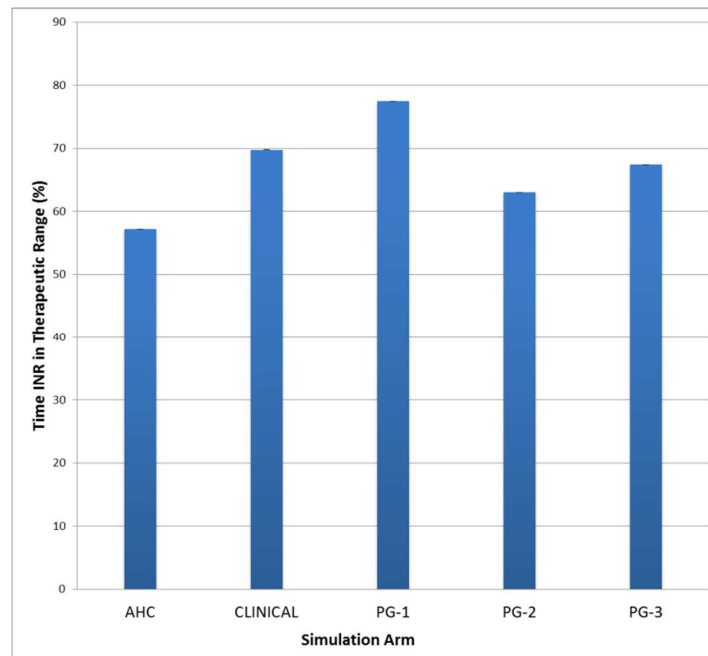
Figure 5.22. TTR (Rosendaal) across the whole clinical avatar population for the 5 simulation arms.

Gender differences are consistent between all five arms. That is, males perform similarly in AHC, Clinical, PG-1, PG-2 and PG-3 when compared to females in those same arms. Interestingly, when the genders are also segregated by race as shown in Figure 5.23, gender difference remains consistent between white and African-American subpopulations. All five arms demonstrate significant difference between African-American males versus white males and African-American females versus white females (P<0.05). For all four subpopulations, PG-1 demonstrated superior predicted percent TTR. AHC arm demonstrates the largest difference between race and gender groups while PG-3 demonstrates the smallest difference between African-

American and white racial groups. African-American males and African-American females demonstrate no significant difference between PG-3 and AHC dosing arms.
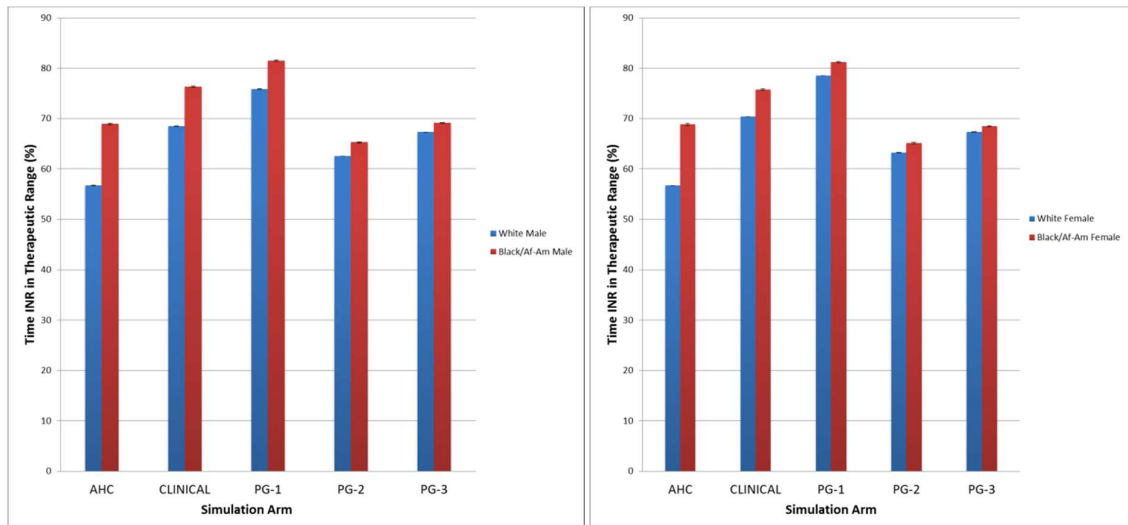
Figure 5.23. TTR (Rosendaal) across the whole clinical avatar population for the 5 simulation arms by gender and race.

Results for clinical avatars when segregated by CYP2C9 genotype are shown in Figure 5.24. Poor and intermediate metabolizers demonstrate significantly lower TTR when compared with the wild type *1/*1 CYP2C9 extensive metabolizer. All three PG-based subpopulations produced superior mean TTR's within the PG-1 arm. Similarly, all three produced inferior mean predicted TTR's in the AHC arm. For extensive metabolizers, which are a majority of the AHC population, the clinical arm produced the second highest mean predicted TTR, followed by the PG-3 arm and the PG-2 arm. For intermediate and poor metabolizers, the PG-3 arm produced the second highest predicted mean TTR's and the clinical and PG-2 arm produced similar TTR results.
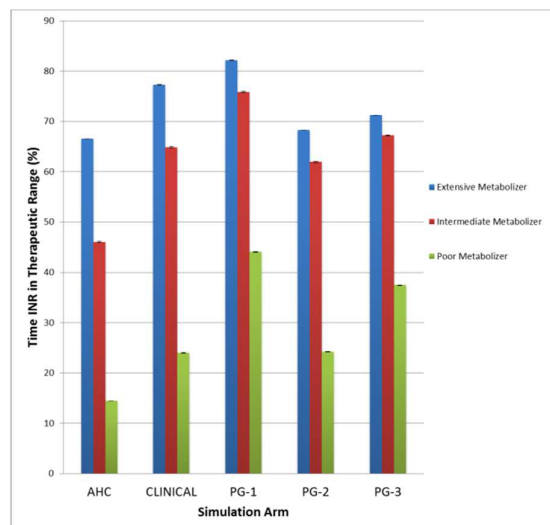
Figure 5.24. TTR (Rosendaal) across the whole clinical avatar population for the 5 simulation arms by CYP2C9 genotype. Extensive Metabolizer: *1/*1, Intermediate Metabolizer: *1/*2, *1/*3, *2/*3, Poor Metabolizer: *2/*2, *3/*3.

In the final figure (Figure 5.25) we look the percent of time above therapeutic range and the percent of time below therapeutic range as calculated by linear interpolation. The results are shown for the whole population as well as two subpopulations white race and African-American subpopulations. The AHC arm produced the highest predicted time above therapeutic range for the whole population and white and African-American subpopulations. While PG-3 arm produced the lowest time above therapeutic range for the whole population and white clinical avatars. While the arms PG-1 and PG-3 produced similar time above therapeutic range specifically for the African-American subpopulation.

In contrast, the PG-2 and PG-3 produced the highest mean time, greater than 20%, with subtherapeutic INR for the whole population and the white and African-American subpopulations. The AHC arm and the PG-1 arm produced similar predicted times subtherapeutic range for the white population and for the white population, but African-American demonstrated greater time below

therapeutic range in the PG-1 arm when compared to the AHC arm. The clinical arm produced results that demonstrate slightly greater predicted time below therapeutic range when compared to the PG-1 arm. Therefore, white subpopulations demonstrate the least time subtherapeutic range in AHC and PG-1 arms while African-American subpopulation demonstrates the least time in subtherapeutic range specifically in the AHC arm.
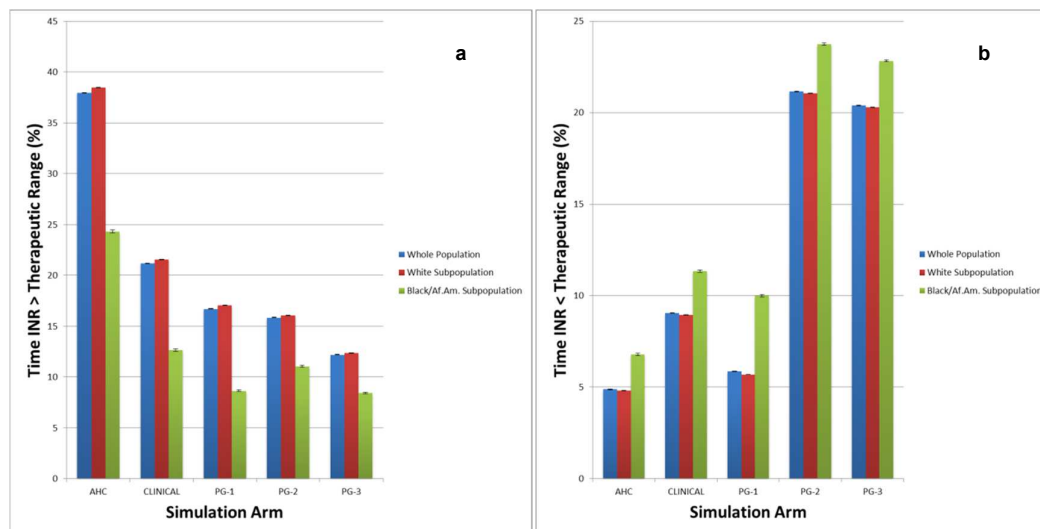
Figure 5.25. (a) Percent Time INR higher than therapeutic range across the five simulation arms by race. (b) Percent Time INR lower than therapeutic range across the five simulation arms by race.

5.4.4.) Conclusion

We have simulated a total of five clinical trial arms, replicating each arm 100 times. We simulated 1.5 million clinical avatars for each arm bringing the total simulated population to 7.5 million. We found that the PG-1 arm produced superior predicted clinical outcomes across the whole AHC population and all relevant subpopulations within the study population. When comparing the top performing arm (i.e., PG-1) to the current Aurora Health Care best practice warfarin dose protocol (i.e., AHC arm), we demonstrate superior TTRs for the whole population and all subpopulations. Therefore, we confirm our primary

hypothesis that there exists several warfarin dosing algorithms that provide a measure of personalization, in this case PG-based and clinically-based, that improved clinical outcomes across the entire AHC population.

The superiority of the PG-1 arm was mirrored in the secondary outcome metrics of predicted time above therapeutic range and predicted time below therapeutic range. The PG-1 arm produced lower predicted time above therapeutic range when compared to the AHC arm for all populations and subpopulations, reducing the risk of intracranial hemorrhage and other forms of bleeding. The PG-1 arm produced similar predicted time below therapeutic range for the white subpopulation, however, the African-American subpopulation did experience significantly greater percentage time below therapeutic range, thereby diminishing its pharmacological effectiveness.

We performed a total of four clinical trial arms that included some degree of warfarin dose personalization (Clinical, PG1, PG-2 and PG-3). We defined personalization as any demographic, clinical or genetic variable applied across a medication dosing protocol. The PG-2 and PG-3 clinical trial arms included patient specific dose personalization at both the initial dose prediction and the adjustment stage of the algorithm. In contrast, PG-1 and Clinical arms only included warfarin dose personalization at the initiation stage of warfarin therapy. As noted above PG-1 produced superior clinical outcome metrics for all subpopulations within the AHC study population. Additionally, the clinical arm produced non-inferior clinical outcome metrics to the PG-2 arm and PG-3 arm for the whole AHC population. However, PG-based subpopulations did demonstrate superior outcomes with the PG-2 and PG-3 arms.

Nevertheless we must reject our second hypothesis that predicted clinical outcomes improve proportionally to the degree of dose personalization across the entire AHC population. While personalization in general provides superior outcomes when compared to a "one-size fits all" approach for the AHC best-practice warfarin management, greater inclusion of personal characteristics within the dosing algorithm does not improve clinical outcome metrics across the whole AHC population for primary outcome metrics.

5.5.) References:

Anderson, J. L., Horne, B. D., Stevens, S. M., Grove, A. S., Barton, S., Nicholas, Z. P., et al. (2007). Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. Circulation, 116(22), 2563-2570.

Anderson, J. L., Horne, B. D., Stevens, S. M., Woller, S. C., Samuelson, K. M., Mansfield, J. W., et al. (2012). A randomized and clinical effectiveness trial comparing two pharmacogenetic algorithms and standard care for individualizing warfarin dosing (CoumaGen-II). Circulation, 125(16), 1997-2005.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics, 77(2), 81-97.

Clyde, M. A., & Lee, H. K. (2001). Bagging and the Bayesian bootstrap. In Artificial Intelligence and Statistics (Vol. 169174).

Fiore, M. C., Novotny, T. E., Pierce, J. P., Hatziandreu, E. J., Patel, K. M., & Davis, R. M. (1989). Trends in cigarette smoking in the United States: the changing influence of gender and race. Jama, 261(1), 49-55.

Fusaro, V. A., Patil, P., Chi, C. L., Contant, C. F., & Tonellato, P. J. (2013). A systems approach to designing effective clinical trials using simulations. Circulation, 127(4), 517-526.

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. Neural Computing and Applications, 19, 263-282.

Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Statistics in medicine, 26(29), 5320-5334.

Kovacs, M. J., Rodger, M., Anderson, D. R., Morrow, B., Kells, G., Kovacs, J., et al. (2003). Comparison of 10-mg and 5-mg warfarin initiation nomograms together with low-molecular-weight heparin for outpatient treatment of acute venous thromboembolism. A randomized, double-blind, controlled trial. Annals of internal medicine, 138(9), 714-719.

Kuskowska-Wolk, A., Bergström, R., & Boström, G. (1992). Relationship between questionnaire data and medical records of height, weight and body

mass index. International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity, 16(1), 1-9.

Marot, A., Morelle, J., Chouinard, V. A., Jadoul, M., Lambert, M., & Demoulin, N. (2011). Concomitant use of simvastatin and amiodarone resulting in severe rhabdomyolysis: a case report and review of the literature. Acta Clinica Belgica, 66(2), 134-136.

Pirmohamed, M., Burnside, G., Eriksson, N., Jorgensen, A. L., Toh, C. H., Nicholson, T., et al. (2013). A randomized trial of genotype-guided dosing of warfarin. The New England journal of medicine, 369(24), 2294-2303.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved August 1, 2014, from http://www.R-project.org/

Rosendaal, F., Cannegieter, S., Van Der Meer, F., & Briet, E. (1993). A method to determine the optimal intensity of oral anticoagulant therapy. Thrombosis and haemostasis, 69(3), 236-239.

Scheines, R., Spirtes, P., Glymour, C., Richardson, T., & Meek, C. (1998). The TETRAD Project : Constraint Based Aids to Model Specification. Multivariate Behavioral Research, Vol. 33, N. 1, 65-118, & "Reply to Commentary," same issue, 165-180.

Scott, S. A., Khasawneh, R., Peter, I., Kornreich, R., & Desnick, R. J. (2010). Combined CYP2C9, VKORC1 and CYP4F2 frequencies among racial and ethnic groups. Pharmacogenomics, 11(6), 781-791.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search. 2nd edition. New York, NY: MIT Press.

University of Wisconsin-Milwaukee. (2014). High Performance Computing Research Cluster Avi. Retrieved July 17, 2014 from www.uwm.edu/hpc/

Wilson, S. E., Costantini, L., & Crowther, M. A. (2007). Paper-based dosing algorithms for maintenance of warfarin anticoagulation. Journal of thrombosis and thrombolysis, 23(3), 195-198.

Chapter 6: Secondary use of electronic medical records to enhance *in silico* comparative effectiveness research: An application to anticoagulation health disparity

6.1.) Background and Significance

The Patient Protection and Affordable Care Act (PPACA, 2010) established Patient-Centered Comparative Effectiveness Research (PC-CER) as a US national medical research priority. Under PPACA, the Patient-Centered Outcomes Research Institute (PCORI, 2014) was charged to support research that investigates effectiveness and outcomes of health care treatments, medical services, and clinical care in subpopulations (Methodology Committee of PCORI, 2012). Subpopulations may be defined by race, ethnicity, gender, age, and medical and physiologic inclusion criteria such as disease, comorbidities, genotype or cancer molecular subtype. Current PCORI research focus includes secondary use of electronic medical records (EMR) to build clinical research databases (Selby, 2012) to conduct a spectrum of PC-CER studies including health care disparity in part by seeking evidence of treatments' effectiveness across various populations. Socioeconomic status (SES) is one of the most powerful drivers of population-level health outcomes and lower SES is consistently associated with poorer health outcomes (Adler, 1994a, 2008b). Consequently, one such CER area is health care outcome disparity observed across SES-based on education and income (PCORI, 2014). Another PC-CER area of immense interest to biomedical scientists and medical geneticists is the use and value of genetic tests, data and information intended to improve public health. No studies to date have tested if genetics can reduce SES-based health disparity outcome. Warfarin treatment effectiveness studies have potentially high health SES-based disparity impact when comparisons include genetic-based against other "best practice" means of achieving therapeutic dosing. Warfarin's

therapeutic dosing is complicated by a narrow therapeutic index and large interindividual dose variability (up to 20-fold depending on genotypes, physiology, and compliance (e.g. Anderson, 2007; Momary, 2007; Wu, 2007)). Compared to controlled clinical trial studies, optimization of warfarin dosing in the clinic using a "best" practice protocol is challenging due to variances in treatment monitoring (Mega, 2014) and must be balanced to prevent thromboembolism while avoiding overdosing that increases risk to bleeding events (Flaherty, 2007; Lip, 2011).

PC-CER human subject studies to improve anticoagulation outcomes by optimal selection of warfarin dosing protocols is impractical (too large, complex and costly) if designed to test all or even many of the published pharmacogenetic-based (PGx) and non-PGx dosing algorithms applied to important (and numerous) CER subpopulations. We have created and validated a pharmacogenetic clinical trial modeling and simulation platform (Fusaro, 2013) to conduct *in silico* complex CER simulations to test PGx treatment protocols against key patient subpopulations with a goal to predict improved treatment outcomes. In this study, we extend the application of the platform to include EMR data of a representative large US healthcare system with diverse population to conduct a CER study between warfarin treatment protocols. In addition to typical CER study design factors, we profile the study population by SES and indicate how outcomes may be affected by educational and income status as experienced in Milwaukee.

6.2.) Objective

We create an anticoagulation patient EMR database from patients treated by warfarin at Aurora Health Care system (AHC), a Milwaukee based network,

where an institution-wide standardized warfarin best practice treatment protocol has been in effect for the last decade. Thereafter, we use the database to identify SES outcome disparity and as input to our simulation platform that models patient subpopulations, individual patient treatment and outcomes, simulates use of multiple anticoagulation protocols, predicts patient and population outcomes, and tests those predictions against patient data extracted from the database. We then execute an *in silico* study that tests four warfarin dosing protocols against two (simulated) EMR-based SES subpopulations to predict if any of the four protocols reduces the outcomes disparities. This PC-CER study demonstrates how large comprehensive EMRs covering diverse patient populations, coupled with novel modeling and computational simulations, provides opportunity to conduct and in part validate, *in silico* CER in diverse populations.

We first provide a general description of the *in silico* approach, then describe and apply the methods used to test differences between PGx versus non-PGx anticoagulation treatment in a diverse Milwaukee population, and finally, demonstrate how the results can be used to demonstrate treatment outcomes such as "percent Time in Therapeutic Range" (TTR) of International Normalized Ratio (INR) and frequency of INR tests (Gouin-Thibault, 2010; Koertke, 2003; Sawicki, 1999; White, 1989; Horstkotte, 1998) and validation stratified by various factors including SES.

6.3.) Materials and Methods

The *in silico* PC-CER approach starts with extraction of EMR patient data pertinent to the objective of the study. In this study, we extract and transform

anticoagulation patient records from AHC over the period 2002-2011 (Figure

6.1 Panel A. AHC EMR).
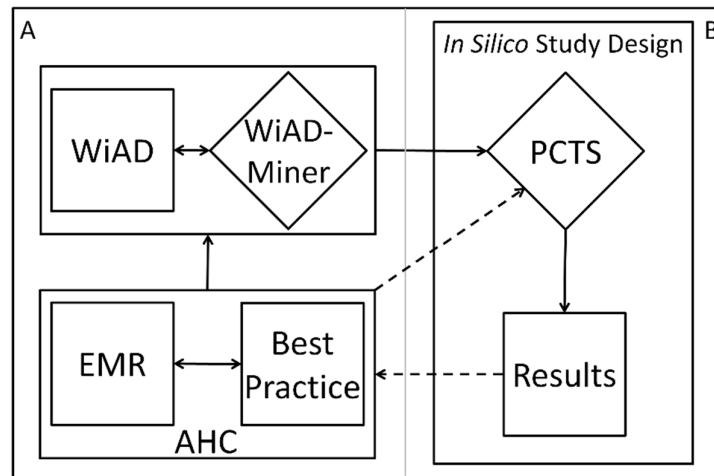
Figure 6.1. Our two component *in silico* PC-CER approach consists of a process for the secondary use of EMR data (Panel A) and a study design and simulation platform (Panel B). The iterative use of the two components can be applied to comparative effectiveness and healthcare disparity research.

Next, we uploaded the data into a minable database (Wisconsin

Anticoagulation Database, WiAD, Figure 6.1, Panel A.) and created a data

mining application (WiAD-Miner). The *in silico* approach then requires a study

design complementary to the objective of the PC-CER and uses our

previously published Pharmacogenetic Clinical Trial Simulation (PCTS)

platform (Figure 6.1, Panel B), (Fusaro, 2013) and an iterative PC-CER

modeling workflow that couples the strength of the EMR database with the

simulation platform (Figure 6.1.). Testing and validating proposed improved

treatment protocols against best practice are represented by dotted lines in

Figure 6.1. Herein, we present the methods for each component of the *in*

silico PC-CER approach when applied to test if genetics can reduce SES-based outcome disparity in anticoagulation therapy.

6.3.1.) Secondary Use of Large Diverse Healthcare EMR

AHC is the largest health care system in Wisconsin serving approximately 1.2 million unique patients each year through 7.8 million patient encounters per year. AHC's EMR is the most comprehensive (by size, type of health care and period of time) digitized health care resource of Southeast Wisconsin's population capturing urban, suburban, and rural constituents of all racial, ethnic, and socioeconomic backgrounds. This resource provides a unique opportunity to capture data and information vital to conducting retrospective and predictive PC-CER studies. However, the same factors that make AHC's EMR extremely valuable (size, scope and longitudinal extent) to PC-CER objectives in general, create great difficulty in capturing the targeted data important to a particular PC-CER study. Consequently, we have created a modular and replicable method to identify, extract, transform and load process and tools to mine EMRs and produce a highly enriched PC-CER knowledgebase that provides both input to and validation of our CER studies.

6.3.1.1.) EMR Extraction, Transformation, and Loading

The AHC EMR was mined to extract all patients with evidence of prescription of: Coumadin (Warfarin), Heparin, Ticlopidine (Ticlid), Clopidogrel (Plavix), Dipyridamole (Persantine), Abciximab (ReoPro), Eptifibatide (Integrilin), Tirofiban (Aggrastat), or Dabigatran (Pradaxa) over the period of 2002 to 2011. Patient data was de-identified per IRB approval (allowing zipcode) by an AHC honest broker before distribution to the research team. Longitudinal data records of 157,450 patients including: gender, race, height, weight, age,

day of visit, patient's zipcode, patient's city, provider's zipcode, smoking

status, INR, medications received (day, dose, frequency), interacting

medications (Amiodarone, Simvastatin, Fluvastatin, Lovastatin, Atrovastatin,

Rosuvastatin, Pravastatin, Aspirin), medication indications (by ICD-9 codes:

Orthopedic surgery-hip or knee, Deep Vein Thrombosis (DVT), Pulmonary

embolism, Atrial fibrillation, Atrial flutter, Atrial fibrillation and flutter, Stroke,

Heart valve replacement) and comorbidities (by ICD-9 codes: DVT,

Pulmonary embolism, Stroke, Myocardial infarction, Bleeding). Subsequent

data cleaning, quality control and quality assurance include an iterative

process of data parsing to detect irregularities; statistical analysis designed to

test population-wide distributions and possible biases; refinement of inclusion

and extraction data mining codes to address irregularities, possible missing

data and detected biases; and ultimately, data transformation to produce a

cohesive set of records capturing all available medical records in a consistent

format following Weiskopf (2013). Representative of our process was the

complex method to produce consistent primary and secondary anticoagulation

outcome metrics such as longitudinal metric TTR in targeted range and

frequency of INR values. INR frequency is required to assess outcome

metrics from EMR data since patients seen in the best practice clinical setting

typically do not experience the same frequency of INR monitoring and

corresponding dose adjustment as experienced for the controlled clinical trial

setting (Mega, 2014).

All extracted INR values and frequency were tested against physiological,

treatment and compliance consistency criteria such as exclusion criteria: (a)

INR values ≤ 1.2 since such values are likely unrelated to warfarin therapy

(White, 1989), (b) INR values ≥ 10 consistent with the upper limit of AHC lab's reference range and (c) problematic INR values during fixed 5mg/day warfarin exposure periods (Sagreiya, 2010). We also profiled patients by defined treatment periods (e.g. fixed time or medical procedure periods) during which TTR calculations are likely well defined. For example, the warfarin exposure period was defined by a combination of time period and frequency of INR values (at least two INR values in a 90 day window). For those individuals "profiled" by the defined treatment period, the TTR was calculated using the linear interpolation method of Rosendaal (1993). This method assumes a linear relationship between two consecutive INR results, assigns an INR value to each day between successive observed INR values, and determines the proportion of time for which the INR is below, within or above the therapeutic range (i.e., time in which patient INR values were between 2 and 3). Then, the individual warfarin exposure period TTRs and mean TTR ($\overline{TTR}$) was calculated for each patient.

6.3.2.) WiAD and WiAD-Miner

After rigorous data extraction, quality control and transformation, the de-identified patient record "cleaned" data (157,450 records reference above), tagged profiles and related metadata were loaded into WiAD. The WiAD patient subpopulation of AHC includes 49.65% female and 50.35% male with mean age of 67.99 yo (female) and 65.22 yo (male). WiAD patients are geographically distributed across all 72 counties of Wisconsin and WiAD's racial distribution is consistent with that of the state. 47.8%, 10.4% and 1.9% of WiAD patients have evidence of only 1, 2, or 3 medications respectively. In addition to the AHC patient data, WiAD includes complementary data such as

Wisconsin population statistics, demographics, and US census data to expand the "knowledgebase" and provide more robust information for the subsequent simulations and predictions. An interactive data profiling and population "segmentation" tool (WiAD-Miner) includes all the profiling (e.g. warfarin only exposure patients), outcome metric (e.g. $\overline{TTR}$ based on warfarin exposure period) and related data analysis functions described above and developed in R (R Core Team, 2013). WiAD-Miner includes a cohort selection tool to profile and identify patient subpopulations by any or a combination of patients' characteristics including gender, age group, race, patient's residence zipcode, provider's zipcode, medication, medication exposure, duration of treatment, number of dose records, frequency of INR values, medication indication, and comorbidities. WiAD-Miner allows adjustment of various parameters such as the medication exposure period definition, triggering a re-profiling and thereafter, re-calculation of outcome metrics.

6.3.3) Pharmacogenetic Clinical Trial Simulator (PCTS)

Our pharmacogenetic clinical trial simulator (Figure 6.1. Panel B) consists of the 5 following adjustable modeling components: 1) A Bayesian network model (BNM) derived from a study population to produce the virtual patient population ("Clinical Avatars") consistent with study population, 2) A module that sets study conditions such as number of subjects, initial dosing, length of study, number of replications, and similar, 3) A circulating medication concentration and INR predictor based on appropriate pharmacokinetic/pharmacodynamics (PK/PD) model (e.g. Hamberg, 2007), 4) A treatment dose algorithm that uses INR or other pertinent physiological

variables and invokes a treatment protocol (see details for warfarin treatment below), and 5) A study healthcare outcome calculator (e.g., TTR). This five-component simulator was validated by significant testing against major pharmacogenetic anticoagulation clinical trials such as CoumaGen-I and –II (Anderson, 2007a, 2012b). The validation included statistically consistent clinical simulations and predictions against the published 90 day, multi-treatment protocol CoumaGen-I results (Fusaro, 2013). All simulations were implemented in R (R Core Team, 2013) and performed on the local high performance computing research cluster Avi (UWM, 2014). The PCTS platform includes the generation of Clinical Avatar populations that mirror the study population's statistical characteristics and are consistent with the EMRs of actual patients. Thereafter, the representative synthetic patient populations are used to conduct replicated clinical simulations testing multiple anticoagulation medication-protocol options.

6.3.4.) Study Population

In this study, WiAD-Miner was used to identify all WiAD patients exposed only to warfarin from 2002-2011 whose records include complete demographic (e.g. race) and geographic data (n=16,900), hereafter this group is called the WiAD warfarin population. For the purpose of calculating treatment outcomes and identification of SES status in this study, the WiAD warfarin population subjectswhose records satisfied the following criteria were selected (a) the inclusion criteria: (1) zipcode in Milwaukee, (2) treatment periods between two successive INR values of 90 days or less and (b) the exclusion criteria: (1) periods of warfarin exposure interruptions (e.g., hospitalization), and (2) 1 week before and 3 weeks after warfarin exposure interruptions. With these

criteria, WiAD-Miner identified 1085 WiAD warfarin population subjects who were then socioeconomically profiled by SES levels across the three "Milwaukee SES zipcode" groups as defined in the 2012 Milwaukee Health Report (Chen, 2012) resulting in a "lower SES" (n=191), "middle SES" (n=716) and "upper SES" (n=178) subpopulations.

6.3.5.) *In Silico* PC-CER Study Design

The PC-CER Study design is based on the comparative effectiveness research aim to detect protocols that improve outcomes and minimize disparity. The PCTS simulation platform requires three specifics: Medication(s) and treatment protocols to be tested; Subpopulation(s) to be studied; and the length of the treatment period and the treatment outcomes. Warfarin dosing protocols typically include three components - an early (1-2 day) warfarin loading dose "initial" protocol followed by a relatively short dose "adjustment" protocol (typically 3-5 days) followed by the therapeutic dose "maintenance" period as needed for treatment. For these simulations, we use four three-component warfarin dosing protocols over a fixed 90 day simulation period: non-PGx CoumaGen-I Standard (Anderson, 2007) PGx CoumaGen-I PG (Anderson, 2008), PGx CoumaGen-II PG-2 Arm (Anderson, 2012) and the PGx Wilson as defined in Fusaro (2013) denoted CG-I STD, CG-I PG and CG-II PG-2, and "Wilson" respectively (Figure 6.2).

Clinical Avatars Study Population: The Milwaukee lower SES and the Milwaukee upper SES subpopulations as defined above were used in this study. A BNM was developed and trained on the WiAD warfarin population. The BNM method is described in Fusaro (2013). The Directed Acyclic Graph (DAG) produced by fitting the WiAD warfarin population to the optimal BNM

was used to produce all clinical avatars used in this study. The BNM's conditional probability table ("CPT") was adjusted to reflect the correlation structure found in the data of the two subpopulations. CPT prior probabilities for the BNM require statistical characterization of the subpopulation including age, gender, weight, height, race, body surface area (calculated from height and weight), tobacco use status, amiodarone use status, and genotypes for CYP2C9 and VKORC1. Statistics of the demographic and clinical information for each of the subpopulations were calculated by WiAD-Miner. The genotype statistics came from a study which determined individual and combined frequencies of important genetic variants associated with warfarin metabolism in several racial and ethnic groups (Scott, 2010). Using the BNM and related R software, we created a sufficient number of clinical avatars (191 x 1000 replicates for lower SES and 178 x 1000 replicates for upper SES) to conduct one thousand parallel simulations of each arm of the study (Figure 6.2).

PC-CER Simulation: For each Clinical Avatar study set (191 lower clinical avatars and 178 upper clinical avatars), a 90 day simulation was performed for each Clinical Avatar using each of the four PGx and non-PG protocols. All results were stored in a structured format representing the 1000 x 90 day study simulations for each clinical avatar in the study subpopulations. Simulation records included: clinical avatar record, simulated INRs and dose values (1 per day for each of 90 days) and calculated TTRs using Rosendaal method (Navathe, 2011). Statistical analysis was applied across the 1000 replicated simulations between the two SES groups.
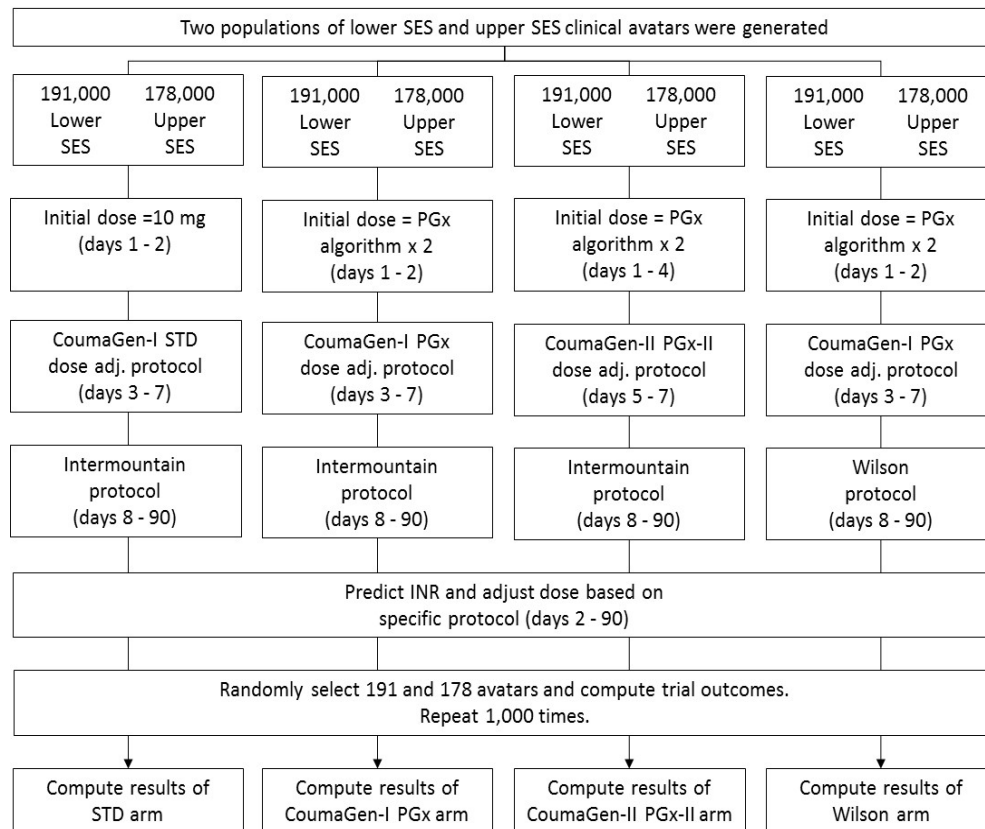
Figure 6.2. The study's design with 4 arms.

6.3.6.) Statistical Analysis

All statistical analyses were conducted with R (R Core Team, 2013).

Statistical significance was set at p<0.05, unless otherwise noted. Single

factor differences between subpopulations were tested using either a

parametric (unpaired t-test, for normal distributions) or a nonparametric test

(Wilcoxon-Mann-Whitney, for non-normal distributions). Comparisons

between the simulated cohorts for $\overline{TTR}$ were made by using one-way ANOVA

tests along with the Tukey post-hoc honestly significant difference (HSD) test

to examine the $\overline{TTR}$ variances across the SES-based subpopulations. Two

way ANOVA test was used to detect possible interactions between

characteristics race and SES on $\overline{TTR}$ for data corresponding to the WiAD's

cohorts. Results are expressed as mean ± standard error of the mean (SEM) unless otherwise specified.

6.4.) Results

6.4.1.) General Characteristics of WiAD Warfarin Cohorts

A total of 369 warfarin patient were identified across the Milwaukee lower SES (n=191) and upper SES (n=178) groups (2nd and 3rd columns of Table 6.1). Mean (SD) age of upper and lower SES patients were 73.74 (14.4) and 61.69 (16.18) respectively. There was no significant difference (p>0.05) in either gender or height but average weight and tobacco use of the lower SES was higher than the upper SES cohort. The racial profile of the lower SES cohort was very different from the upper SES cohort (e.g. 53.9% versus 93.8% white).

6.4.2.) General Characteristics of Clinical Avatar Cohorts

The statistical characteristics of the clinical avatars simulated to represent the study subpopulations, are presented in the 4th and 5th columns of Table 6.1. All demographic and clinical characteristics of the clinical avatar cohorts were statistically the same as the Milwaukee warfarin lower and upper SES study cohorts. Genotype frequencies were matched to those published for populations equivalent to the AHC (Scott, 2010).

6.4.3.) WiAD Warfarin Cohorts' Outcome Metrics

The outcome metrics for this study were $\overline{TTR}$ and number of INR (Figures 6.3 and 6.4). Lower SES cohort had significantly lower $\overline{TTR}$ compared to Upper SES (39.82%±1.9 vs 48.88%±1.83, P<0.05). Lower SES cohort had

Table 6.1. Basic characteristics of the WiAD warfarin and clinical avatars cohorts. (1) The WiAD warfarin cohorts' genotypes were imputed using Scott's distributions, correlations and joint distributions (Scott, 2010).

Characteristics	WiAD Warfarin Cohorts		Clinical Avatar Cohorts	
	Milwaukee Lower SES	Milwaukee Upper SES	Lower SES	Upper SES
Number of Patients	191	178	191	178
Age, year, mean (SD)	61.6 (16.1)	73.74 (14.4)	61.29 (15.9)	73.38 (14.3)
Gender, female %	59.1	56.1	59.2	57.4
Weight, kg, mean (SD)	91.2 (26.2)	84.5 (24.7)	94.4 (25.4)	87.7 (24.6)
Height, in, mean (SD)	66.13 (4.2)	66.25 (4.9)	66.28 (3.7)	65.78 (3.7)
Race, %				
White	53.9	93.8	53.8	93.9
African American	44.5	5.6	44.8	5.4
Asian	1.6	0.6	1.3	0.6
Tobacco use, %	10.9	5.0	10.7	5.3
Amiodarone use, %	12.0	12.9	11.4	13.2
DVT, %	26.1	17.4	29.7	18.4
VKORC1, %	(1)	(1)		
A/A	10.5	16.9	9.7	16.4
G/A	32.5	43.5	33.0	42.1
G/G	56.8	39.4	57.2	41.2
CYP2C9, %	(1)	(1)		
*1/*1	70.4	66.5	71.1	67.4
*1/*2	10.1	14.4	8.6	13.3
*1/*3	6.5	9.0	6.0	8.5
*2/*2	3.6	6.2	3.0	5.4
*2/*3	1.1	1.8	1.0	1.9
*3/*3	0.7	0.2	1.1	0.3

significantly lower number of INR than Upper SES (7.0 vs 13.0, median, respectively, p<0.05). No significant interaction between SES group and race on $\overline{TTR}$ was present (two-way ANOVA, p<0.05). In the absence of interaction with race, SES had significant effect on $\overline{TTR}$.

6.4.4.) Clinical Avatar Cohorts' Outcome Metrics

Figure 6.3 presents the Milwaukee lower (left panel) and upper (right panel) SES clinical avatar cohorts' TTR computed for CG-I STD, CG-I PG, CG-II PG-2 and Wilson in order.

Figure 6.3. Averaged $\overline{TTR}$ for lower and upper SES clinical avatar cohorts by dosing protocol (left and right panel). Capped vertical lines represent standard error of the mean. For both cohorts, the Wilson protocol's TTR was significantly higher than any other protocol (ANOVA, p<0.05).

For the lower SES cohort, all PGx protocols produced significantly higher results than the non-PGx protocol and the Wilson protocol was highest (42.41±2.26 vs 52.04±2.24, 56.53±1.93, 73.84±1.52, ANOVA, p<0.05). The same result was true for the upper SES cohort (38.85±0.74, 52.58±0.53, 52.46±0.44, 75.66±0.49, ANOVA, p<0.05). The averaged $\overline{TTR}$ predicted by the Wilson protocol for the two clinical avatar cohorts were not significantly different. The average 90 day frequency of INR predicted for each protocol was shown in Figure 6.4 (Lower SES, left panel, Upper SES, right panel). For the lower SES cohort, the CG-I STD protocol predicted significantly higher frequency of INR than either CG-I PG or Wilson protocol (10.2±0.4 vs 9.4±0.16 and vs 8.7±0.3, ANOVA, p<0.05). Whereas in the upper SES clinical avatar cohorts, the predicted frequency of INRs were not different.
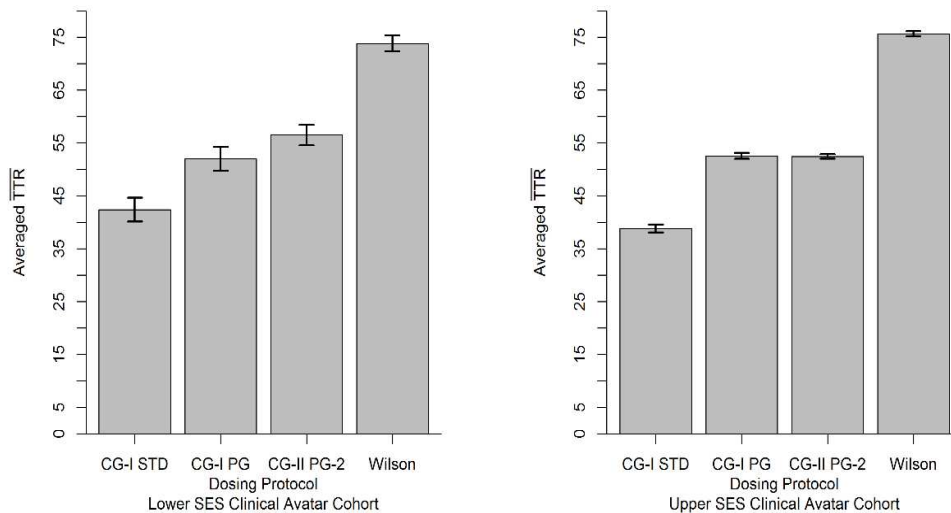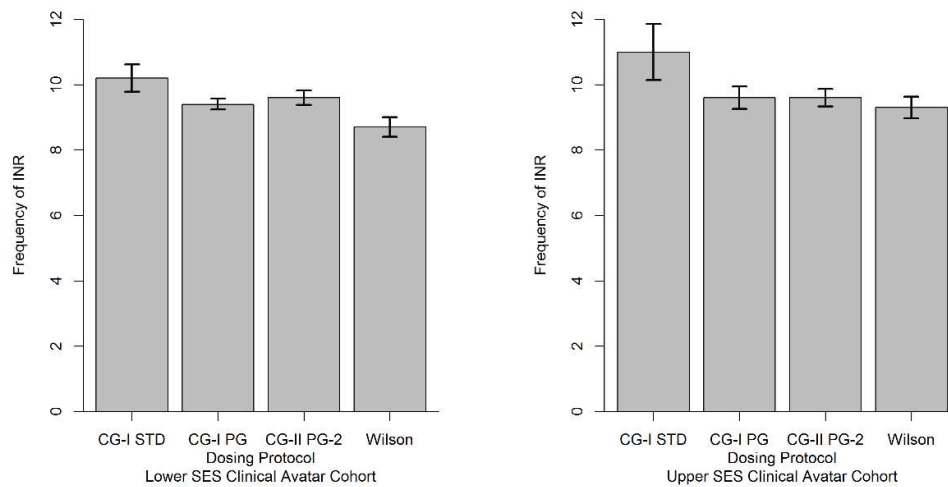
Figure 6.4. Averaged frequency of INRs for lower and upper SES clinical avatar cohorts by dosing protocol (left and right panel). Capped vertical lines represent standard error of the mean. For both cohorts, the Wilson protocol's frequency of INRs was significantly lower than any other protocol (ANOVA, p<0.05).

6.4.5.) Comparison between WiAD Warfarin Cohorts and Clinical Avatar Cohorts

Averaged $\overline{TTR}$ was not significantly different between the Milwaukee lower SES cohort (calculated from EMR data) under current AHC non-PGx best practice protocol when compared to the equivalent clinical avatar cohort under the CG-I STD dosing protocol (39.82%±1.9 vs 42.41±2.26) and was significantly less than the three PGx protocol predictions (39.82%±1.9 vs 52.04±2.24, 56.53±1.93, 73.84±1.52, ANOVA, p<0.05). Averaged $\overline{TTR}$ for the Milwaukee upper SES cohort is significantly more than that predicted for the upper SES clinical avatars using CG-I STD (48.88%±1.83 vs 38.85±0.74) and significantly less than that predicted using the Wilson protocol (48.88%±1.83 vs 75.66±0.49).

6.5.) Discussion

In this study, we present an *in silico* pharmacogenetic anticoagulation PC-CER approach applied to a patient-centered outcomes simulation that produced a prediction that was validated with patient data acquired directly from a large hospital transaction electronic medical record system. We applied this approach to a well-documented health disparity cohort defined by a Milwaukee SES index. Upper and lower SES patients treated by warfarin using the healthcare system's best practice dosing protocol were extracted from the EMR and simulated to predict treatment outcomes under four different warfarin dosing protocols. The approach we used in this study includes three key elements (Figure 6.1.): a) High quality EMR derived data set and data mining environment (WiAD) from which clinical-trial like outcome metrics are extracted; b) Means to conduct complex multi-factorial anticoagulation simulation study designs (PCTS) and predict outcome metrics from multiple treatment protocols; and c) Means to test the predicted outcome metrics against those derived from the EMR (WiAD-Miner).

It has been suggested that meaningful patient-centered outcomes research need high-quality data, including greater clinical detail, longitudinal follow-up, and linkages among data sets (Navathe, 2011). An important result of this work is the dynamic collection of anticoagulation treated patient data and related information encapsulated in WiAD in support of our PC-CER study. WiAD includes over 150,000 Wisconsin resident patients spanning some ten years of anticoagulation diagnosis, treatment and outcome data derived from the Aurora Health Care EMR and all data was processed using a two-step extraction and transformation method including completeness, correctness,

and concordance (Weiskopf, 2013). We used WiAD to select the two

Milwaukee warfarin SES subpopulations and calculate both $\overline{TTR}$ and number

of INRs. The lower SES subpopulation's $\overline{TTR}$ was significantly less than the

upper SES subpopulation's $\overline{TTR}$ beyond any interaction based on race or

ethnicity alone. This result is consistent with the Milwaukee Heath Report

2012, which indicated that dramatic health disparities by SES exist and persist

within Wisconsin's largest city (Chen, 2012). Our results demonstrate that this

disparity is correlated to SES and may be explained by doctor-patient

relationships as demonstrated in previous studies (Schouten, 2006; Bates,

2009).

The 90-day warfarin therapy simulation predicted similar outcome for the

standard non genotype-based dosing protocol (non-PGx CG-I STD) to those

calculated for the lower SES subpopulation using the EMRs. Surprisingly, the

approach also predicted a similar level of averaged $\overline{TTR}$ (38.85%) for the

upper group under the standard protocol as the lower subpopulation even

though the upper subpopulation's EMR-calculated averaged $\overline{TTR}$ was much

higher (48.88%). This result indicates that the actual outcomes of upper SES

group in the healthcare system is much higher than one would predict with

consistent application of the same protocol across all patients independent of

SES. In addition, the predictions suggest that the warfarin genotype-based

Wilson protocol would produce the highest averaged $\overline{TTR}$ across the two SES

subpopulations while requiring a significantly lower frequency of INR values

for the lower SES subpopulation. Consistent with our predictions for a revised

CoumaGen-I trial (Fusaro, 2013). the Wilson protocol produced the best

outcomes of all protocols and simultaneously dramatically reduced the health disparity between the lower and upper SES subpopulations.

6.6.) Conclusion

Multifactorial PC-CER clinical trials designed to test the optimal treatment against multiple subpopulations create a complex and costly paradigm whose results will guide improved healthcare. Part of this complexity can be addressed using judiciously developed and validated mathematical modeling and simulations. Herein, we demonstrated the value of using EMRs to extract clinical-trial like anticoagulation outcome metrics and detect possible disparities within the study population. We applied the approach to a study population that includes individuals who historically experience health disparity to simulate patient outcomes using one non-PGx and three PGx protocols. Our results indicate that the Wilson genotype-based warfarin protocol applied systematically to all patients, improves outcomes overall and reduces the observed health disparity. If validation studies designed to test these predictions prove true, then the optimal warfarin protocol translated into the healthcare setting will improve best practice as suggested in Figure 6.1's dotted lines. The combination of *in silico* studies followed by carefully designed targeted validation studies, suggests a powerful approach to improve healthcare overall and reduce health disparity in particular.

6.7.) References:

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., et al. (1994). Socioeconomic status and health. The challenge of the gradient. The American psychologist, 49(1), 15-24.

Adler, N. E., & Rehkopf, D. H. (2008). U.S. disparities in health: descriptions, causes, and mechanisms. Annual review of public health, 29, 235-252.

Anderson, J. L., Horne, B. D., Stevens, S. M., Grove, A. S., Barton, S., Nicholas, Z. P., et al. (2007). Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. Circulation, 116(22), 2563-2570.

Anderson, J. L., Horne, B. D., Stevens, S. M., Woller, S. C., Samuelson, K. M., Mansfield, J. W., et al. (2012). A randomized and clinical effectiveness trial comparing two pharmacogenetic algorithms and standard care for individualizing warfarin dosing (CoumaGen-II). Circulation, 125(16), 1997-2005.

Bates, T. R., Connaughton, V. M., & Watts, G. F. (2009). Non-adherence to statin therapy: a major challenge for preventive cardiology. Expert opinion on pharmacotherapy, 10(18), 2973-2985.

Chen H-Y, Baumgardner DJ, Frazer DA, et al. (2012). Milwaukee Health Report 2012: Health Disparities in Milwaukee by Socioeconomic Status. Center for Urban Population Health: Milwaukee, WI. Retrieved September 10, 2014 from http://www.cuph.org/mhr/2012-milwaukee-health-report.pdf

Flaherty, M. L., Kissela, B., Woo, D., Kleindorfer, D., Alwell, K., Sekar, P., et al. (2007). The increasing incidence of anticoagulant-associated intracerebral hemorrhage. Neurology, 68(2), 116-121.

Fusaro, V. A., Patil, P., Chi, C. L., Contant, C. F., & Tonellato, P. J. (2013). A systems approach to designing effective clinical trials using simulations. Circulation, 127(4), 517-526.

Gouin-Thibault, I., Levy, C., Pautas, E., Cambus, J. P., Drouet, L., Mahe, I., et al. (2010). Improving anticoagulation control in hospitalized elderly patients on warfarin. Journal of the American Geriatrics Society, 58(2), 242-247.

Hamberg, A. K., Dahl, M. L., Barban, M., Scordo, M. G., Wadelius, M., Pengo, V., et al. (2007). A PK-PD model for predicting the impact of age, CYP2C9, and VKORC1 genotype on individualization of warfarin therapy. Clinical pharmacology and therapeutics, 81(4), 529-538.

Horstkotte, D., Piper, C., & Wiemer, M. (1998). Optimal Frequency of Patient Monitoring and Intensity of Oral Anticoagulation Therapy in Valvular Heart Disease. Journal of thrombosis and thrombolysis, 5 Suppl 1(3), 19-24.

Koertke, H., Minami, K., Boethig, D., Breymann, T., Seifert, D., Wagner, O., et al. (2003). INR self-management permits lower anticoagulation levels after mechanical heart valve replacement. Circulation, 108 Suppl 1, II75-78.

Lip, G. Y., Andreotti, F., Fauchier, L., Huber, K., Hylek, E., Knight, E., et al. (2011). Bleeding risk assessment and management in atrial fibrillation patients. Executive Summary of a Position Document from the European Heart Rhythm Association [EHRA], endorsed by the European Society of Cardiology [ESC] Working Group on Thrombosis. Thrombosis and haemostasis, 106(6), 997-1011.

Mega, J. L., & Giugliano, R. P. (2014). Genotype-guided dosing of warfarin. Clinical chemistry, 60(7), 920-922.

Momary, K. M., Shapiro, N. L., Viana, M. A., Nutescu, E. A., Helgason, C. M., & Cavallari, L. H. (2007). Factors influencing warfarin dose requirements in African-Americans. Pharmacogenomics, 8(11), 1535-1544.

Navathe, A. S., Clancy, C., & Glied, S. (2011). Advancing research data infrastructure for patient-centered outcomes research. JAMA : the journal of the American Medical Association, 306(11), 1254-1255.

Patient-Centered Outcomes Research Institute (PCORI). (2012). National priorities for research and research agenda. Retrieved August 5, 2014, from http://www.pcori.org/content/national-priorities-and-research-agenda

Patient-Centered Outcomes Research Institute (PCORI) Methodology Committee. (2013). The PCORI Methodology Report. Retrieved July 7, 2014, from pcori.org/research-we-support/research-methodology-standards.

PCORI. (2014). Patient-Centered Outcomes Research Institute. Retrieved April 20, 2014, from http://www.pcori.org/research-we-support/pcor/

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved August 1, 2014, from http://www.R-project.org/

Rose, A. J., Hylek, E. M., Ozonoff, A., Ash, A. S., Reisman, J. I., & Berlowitz, D. R. (2010). Patient characteristics associated with oral anticoagulation

control: results of the Veterans AffaiRs Study to Improve Anticoagulation (VARIA). Journal of thrombosis and haemostasis : JTH, 8(10), 2182-2191.

Rosendaal, F. R., Cannegieter, S. C., van der Meer, F. J., & Briet, E. (1993). A method to determine the optimal intensity of oral anticoagulant therapy. Thrombosis and haemostasis, 69(3), 236-239.

Sagreiya, H., & Altman, R. B. (2010). The utility of general purpose versus specialty clinical databases for research: warfarin dose estimation from extracted clinical variables. Journal of biomedical informatics, 43(5), 747-751.

Sawicki, P. T. (1999). A structured teaching and self-management program for patients receiving oral anticoagulation: a randomized controlled trial. Working Group for the Study of Patient Self-Management of Oral Anticoagulation. JAMA : the journal of the American Medical Association, 281(2), 145-150.

Schouten, B. C., & Meeuwesen, L. (2006). Cultural differences in medical communication: a review of the literature. Patient education and counseling, 64(1-3), 21-34.

Scott, S. A., Khasawneh, R., Peter, I., Kornreich, R., & Desnick, R. J. (2010). Combined CYP2C9, VKORC1 and CYP4F2 frequencies among racial and ethnic groups. Pharmacogenomics, 11(6), 781-791.

Selby, J. V., Beal, A. C., & Frank, L. (2012). The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda. JAMA : the journal of the American Medical Association, 307(15), 1583-1584.

The Patient Protection and Affordable Care Act of 2010 (PPACA). (2010). 111–1482010. Retrieved August 10, 2014, from http://www.gpo.gov/fdsys/pkg/BILLS-111hr3590enr/pdf/BILLS-111hr3590enr.pdf

University of Wisconsin-Milwaukee. (2014). High Performance Computing Research Cluster Avi. Retrieved July 17, 2014 from www.uwm.edu/hpc/

Vila, P. M., Swain, G. R., Baumgardner, D. J., Halsmer, S. E., Remington, P. L., & Cisler, R. A. (2007). Health disparities in Milwaukee by socioeconomic status. WMJ : official publication of the State Medical Society of Wisconsin, 106(7), 366-372.

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association : JAMIA, 20(1), 144-151.

White, R. H., McCurdy, S. A., von Marensdorff, H., Woodruff, D. E., Jr., & Leftgoff, L. (1989). Home prothrombin time monitoring after the initiation of warfarin therapy. A randomized, prospective study. Annals of internal medicine, 111(9), 730-737.

Wu, A. H. (2007). Use of genetic and nongenetic factors in warfarin dosing algorithms. Pharmacogenomics, 8(7), 851-861.

Chapter 7: Conclusions and Future Efforts using *In Silico* PC-CER
Translational Research Framework

7.1.) Conclusion
7.2.) Limitations
7.3.) Future Work
7.4.) References

7.1.) Conclusion

The concept of "translational research" is relatively new. In a commentary in JAMA, Woolf has pointed out that "translational research means different things to different people" (Woolf, 2008). The newly established NIH's National Center for Advancing Translational Sciences has defined "Translation" as "the process of turning observations in the laboratory and clinic into interventions that improve the health of individuals and the public - from diagnostics and therapeutics to medical procedures and behavioral changes" (NCATS, 2014). According to their definition, translational science is "the field of investigation focused on understanding the scientific and operational principles underlying each step of the translational process" (NCATS, 2014). As such, translational research models are designed to "translate" discoveries identified through basic science studies (whether gained in animal models of human disease or through human studies), to knowledge concerning the potential value of the discovery's application in medicine.

Translational research in genomics, specifically pharmacogenomics as one of the first clinical applications of the new genomic era, aims to move promising genomic applications to clinical and public health practice for population health benefit (Cleeren, 2011). Despite the demonstrable benefits of many new genomic discoveries, there have been gaps between the explosive growth in scientific discovery and technology and the implementation of this new knowledge. It is widely recognized that the current translational process is slow, very expensive and often results in an incomplete transfer of research findings into practice, and consequently failure of comparative effectiveness

studies used to translate the findings into substantial changes in patient care and health disparities (Khoury, 2007). A study aiming to evaluate the predictors of and time taken for the translation of highly promising basic research into clinical applications, over a 15-year period, showed that only about 5% of the basic science findings were licensed for clinical use and only 1% were extensively used for the licensed indications (Contopoulos-Ioannidis, 2003).

Although a large number of translational research models have been developed over the time, however, no existing model has adequately addressed the pressing need to create a process and pragmatic approach that will cover the ever expanding collection of high-throughput, individualized data that is generated by ever advancing technology. In short, the deluge of big data especially in the light of expanding electronic medical record systems (EMRs) and genomic era has overwhelmed the antiquated models and processes designed to translate important and growing data and evidence to the healthcare setting. These models are further weakened when considering the important area of patient-centered, comparative effectiveness research and the potential disparity of outcomes when coarse applications are applied to diverse populations. Given these facts and the need to address many complex real-world healthcare questions in short periods of time, it seems that alternative research designs and approaches should be considered in translational research.

Taking into consideration these facts, in this dissertation, I have proposed an iterative and bidirectional agile translational research model enhanced with an *in silico* knowledge synthesis model (*iS*-TR) to facilitate pharmacogenomic

patient-centered comparative effectiveness research (PC-CER) studies. I have hypothesized that retrospective EMR analysis and subsequent mathematical modeling and simulation predictions (a) may facilitate and accelerate the process of generating and translating pharmacogenomic knowledge, (b) may be applied to determine comparative effectiveness of anticoagulation treatment plan(s) tailored to well defined target populations, and (c) may result in a decrease in overall adverse risk and improve individual and population outcomes.

To test the hypotheses, I have developed an *In Silico* PC-CER Approach for warfarin pharmacogenomics knowledge. The two-component *In Silico* PC-CER Approach consists of a process for the secondary use of "Big EMR Data" resulting in a unique anticoagulation database, Wisconsin Anticoagulation Database ("WiAD"), and also a study design and clinical trial simulation platform. Once the simulation platform was validated by replicating and reproducing the results of two major warfarin pharmacogenomic clinical trials of CoumaGen-I and II (Anderson, 2007a, 2012b), the Approach was applied (a) to predict optimal anticoagulation treatment plan for the Aurora Health Care's large heterogeneous patient population, and (b) to an anticoagulation therapy outcomes disparity in City of Milwaukee recognized as the most segregated metropolitan area in the country with a significant SES-based health disparity.

The studies' results have demonstrated that the *In Silico* PC-CER Approach taking advantage of retrospective EMR analysis and subsequent mathematical modeling and simulation prediction could facilitate and accelerate the process of generating and translating pharmacogenomic

knowledge on comparative effectiveness of anticoagulation treatment plan(s) tailored to well defined target populations (leading to a decrease in overall adverse risk and improve individual and population outcomes). Accordingly, we have concluded that the combination of *in silico* studies followed by carefully designed targeted validation studies would be a powerful approach to conduct PC-CER studies, improve healthcare overall and reduce health disparity in particular.

7.2.) Limitations

The In Silico PC-CER Approach provides many opportunities and we have demonstrated that it could be used as an effective approach to improve and facilitate translational research. However, given the complexity of the anticoagulation therapy in real clinical settings especially for warfarin, transfer of knowledge and evidence produced by the Approach to the real world clinical setting requires clinical judgment of the healthcare providers. Even though the Approach provides a great opportunity to conduct comparative effectiveness studies on heterogeneous study populations; however, real world clinical conditions will obviously vary and do not perfectly reflect the content or performance of the Approach. For instance, several foods and herbal supplements can interact with warfarin and affect its effectiveness. The current Approach does not include modules to take into consideration this kind of factors in its predictive outcomes.

7.3.) Future Work

In next few months, I will focus on improving the *In Silico* PC-CER Approach. One important step will be to expand and enrich WiAD by extracting and including more clinical information of the subjects allowing us to take

advantage of other predictive models such as CHA_2DS_2-VASc for risk of ischemic stroke (Lip, 2010) and HAS-BLED for risk of bleeding (Pisters, 2010). The other task will be to design and conduct some validation studies of the predicted subpopulation anticoagulation treatment plans produced by the *In Silico* PC-CER Approach.

I am also interested in applying the *In Silico* PC-CER Approach to other fields such as cancer recurrence and progression. The Approach also has great potential to be used as the basis of a Decision Support System for clinicians.

7.4.) References

Anderson, J. L., Horne, B. D., Stevens, S. M., Grove, A. S., Barton, S., Nicholas, Z. P., et al. (2007). Randomized trial of genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. Circulation, 116(22), 2563-2570.

Anderson, J. L., Horne, B. D., Stevens, S. M., Woller, S. C., Samuelson, K. M., Mansfield, J. W., et al. (2012). A randomized and clinical effectiveness trial comparing two pharmacogenetic algorithms and standard care for individualizing warfarin dosing (CoumaGen-II). Circulation, 125(16), 1997-2005.

Cleeren, E., Van der Heyden, J., Brand, A., & Van Oyen, H. (2011). Public health in the genomic era: will Public Health Genomics contribute to major changes in the prevention of common diseases? Archives of public health = Archives belges de sante publique, 69(1), 8.

Contopoulos-Ioannidis, D. G., Ntzani, E., & Ioannidis, J. P. (2003). Translation of highly promising basic science research into clinical applications. The American journal of medicine, 114(6), 477-484.

Khoury, M. J., Gwinn, M., Yoon, P. W., Dowling, N., Moore, C. A., & Bradley, L. (2007). The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?. Genetics in medicine, 9(10), 665-674.

Lip, G. Y., Nieuwlaat, R., Pisters, R., Lane, D. A., & Crijns, H. J. (2010). Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *CHEST Journal*, *137*(2), 263-272.

National Center for Advancing Translational Sciences (NCATS). (2014). Retrieved December 1, 2014 from http://www.ncats.nih.gov/

Pisters, R., Lane, D. A., Nieuwlaat, R., de Vos, C. B., Crijns, H. J., & Lip, G. Y. (2010). A Novel User-Friendly Score (HAS-BLED) To Assess 1-Year Risk of Major Bleeding in Patients With Atrial FibrillationThe Euro Heart Survey. CHEST Journal, 138(5), 1093-1100.

Woolf, S. H. (2008). The meaning of translational research and why it matters.*JAMA*, *299*(2), 211-213.

CURRICULUM VITAE

Name   Kourosh Ravvaz
Address   Laboratory for Public Health Informatics and Genomics, University of Wisconsin Milwaukee, 3209 N. Maryland Ave., Milwaukee, WI 53211
Email   kravvaz@uwm.edu

Personal Statement

I have practiced as a physician at ER and conducted extensive public health initiatives and realized some years ago that the future of healthcare outcome improvements would be heavily dependent on informatics technology. Consequently, I pursued and have received my PhD in Biomedical and Health Informatics. As a member of Laboratory for Public Health Informatics and Genomics in last few years, I worked on several projects designed to improve and understand the use of genetics in healthcare outcomes. My PhD dissertation was based on a personalized medicine and patient-centered outcomes research objective (improved INR management using optimized warfarin dosing) achieved by secondary use of EMR data (157,000 Aurora Health Care patients), and a predictive modeling and simulation method designed to test the optimal dosing protocol to achieve improved INR across a diverse population. I seek to use my experience and background and similar approach and methods developed in my dissertation and available from other's work to improve healthcare outcomes in large, diverse populations.

Education

Doctor of Medicine, M.D.
Iran University of Medical Sciences, 1998

Medical Doctoral Thesis
"A Prevalence Study on Physical Diseases among Schizophrenic Patients Admitted to Major Psychiatric Hospitals in Tehran, Iran."
This project aimed to determine the epidemiological characteristics of physical diseases among the schizophrenic patients admitted to specialized psychiatric hospitals in Tehran in a period of time and then in follow-up projects to study the burden of the diseases and their effects on the patients' quality of life.

Master of Public Health, MPH
Tehran University of Medical Sciences, 2005

MPH Thesis
"An Ecological Study on the Relationship between Health Determinants and Health Status Indicators in Iran."
The objective of this study was to address the relationship between socioeconomic and environmental health determinants and the health status indicators at the provincial and national levels. The determinants and indicators included children aged under-5 mortality rate, children aged under-5 stunted, maternal mortality rate, per capita total expenditure on health, children and adult malnutrition, life expectancy at birth, immunization coverage rate, utilization of health services, health workforce, and population using improved water and sanitation.

PhD, Biomedical and Health Informatics
University of Wisconsin-Milwaukee, 2015

PhD Dissertation
"Facilitating and Enhancing Biomedical Knowledge Translation: An *in silico* approach to patient–centered pharmacogenomic outcomes"

The goal of this project is to develop and test a patient-centered comparative effectiveness approach to simulate, predict and then conduct validation of a method that accelerates meaningful, timely and efficient translation of genetic knowledge related to improving anticoagulation and anticlotting treatments. We evaluate their impact on public health, and assess their applicability and effectiveness in the broader US mixed patient population, adjusting the treatments for minority special issues (such as genetic differences). Part of this work includes a project established with Aurora Health Care as the largest healthcare provider in Wisconsin. Using Aurora's electronic medical records (EMR), it demonstrates that modeling and simulation of pharmacogenetic clinical trials based on actual EMR patient data can provide predictive evidence useful in detecting "optimal" anticoagulation/anticlotting dosing protocols that reduce adverse drug responses, improve overall patient outcomes, and reduce health disparities in the statewide Aurora patient population.

*** This project is going to fulfill the objectives of a Research Growth Initiative (RGI) proposal funded by the University of Wisconsin-Milwaukee (~$100,000).***

Professional Experience
American Medical Informatics Association (AMIA)
- Member, Steering Committee, Implementation and Optimization Forum, 2012-present
- Organizer and Chair, Education and Reference Team, Implementation and Optimization Forum, 2012-present

University of Wisconsin-Milwaukee
Biomedical and Health Informatics
- Biomedical and Health Informatics Scientist, Laboratory for Public Health Informatics and Genomics, 2011-present
  The focus has been on research projects addressing approaches and methods of biomedical knowledge translation continuum from discovery to the dissemination and adoption of effective interventions.
- Biomedical and Health Informatics Scientist, Collaborative Intelligent Health Information Systems Initiative (CIHISI), National Science Foundation, Partnerships for Innovation-PFI program, 2007-2009
  Mapped nursing concepts to standard ontologies (ICNP and SNOMED-CT) in support of evidence-based nursing initiative. Thus mapped, NLP methods were then used to identify standardized concepts in nursing free-text notes thus creating a "translational method" to determine evidence-based rules that were then used in developing nursing decision support systems.
- Biomedical and Health Informatics Scientist, Using Evidence-Based Nursing Practices and EHR Decision Support to Reduce Fall-Related Patient Injuries in Acute Care, The Agency for Healthcare Research and Quality, 2009
  Focused on nursing decision support systems to prevent fall-related patient injuries in acute care by designing and building a metadata repository following the standard ISO 11179 implemented in USHIK-AHRQ's metadata registry.
- Biomedical and Health Informatics Scientist, Developing a Measure of the Accuracy of the Translation of Research into Practice, The Agency for Healthcare Research and Quality, 2009
  Developed quantifiable measures for accuracy of translation comparing hierarchical meta-data schemas or data dictionaries corresponding to each step (or system) in the translation using existing ontology similarity measures or metadata schema matching algorithms.

Teaching
- Instructor, Anatomy and Physiology Course, Department of Biological Sciences, Fall 2009
- Instructor, Introduction to Health Care Informatics Course, Fall 2010, Fall 2011
- Instructor, Legal, Ethical and Social Issues in Healthcare Informatics Course, Fall 2010, Fall 2011
- Instructor, Public Health Informatics and Genomics (PHIG) Course, Spring 2012
- Instructor, Leadership for Healthcare Professionals Course, Fall 2014

Tehran University of Medical Sciences
Administration
- Chief of Staff to the Chancellor, 2000-2005
- University Council Secretary, 2000-2005
- Member of the Campus Academic and Physical Master Planning Executive Leadership Team, 2000-2005

Medical Practice
- Ziayian General Hospital, 2000-2003
- Sina Teaching Hospital, 2003-2006
    Head of the emergency ward (ER)

Editorial Activities
- Reviewer, AMIA Annual Conference, American Medical Informatics Association, 2014
- Reviewer, Journal of Interprofessional Care, 2014
- Reviewer, The 14th World Congress on Medical and Health Informatics (Medinfo 2013),
    Copenhagen, Denmark, 2012-2013
- Reviewer, Journal of Applied Clinical Informatics, Official eJournal of IMIA (International Medical Informatics Association) and AMDIS (Association of Medical Directors of Information Systems), 2012-present

Research Projects
- Developing Metadata Registries across the Roswell Park's Data Repositories, University of Wisconsin-Milwaukee and Roswell Park Cancer Institute, NY, 2014
- Reconciliation of the Terminology Repositories Used on the User Interface and Corresponding Manuals for 32 GE Healthcare's Diagnostic Cardiology Devices, University of Wisconsin-Milwaukee and GE Healthcare, 2014
- Guiding Warfarin Clinical Trial Design Using Pharmacogenetic Simulations, University of Wisconsin-Milwaukee, Harvard Medical School and Aurora Health Care, 2012-2014
- Predicting Clinical Validity of Bladder Cancer Nomograms, University of Wisconsin-Milwaukee and University of Wisconsin-Madison, 2012-2013
- Biomedical Informatic Analysis of the RNA of NF1 Associated Nerve Sheath Tumors, University of Wisconsin-Milwaukee, University of Alabama at Birmingham, and Harvard Medical School, 2011-2012
- Population-wide Analysis of the Likely Benefits of Genetic Testing in Narrow-Therapeutic Window Drugs. University of Wisconsin-Milwaukee. 2010-2012
- Quantitative Assessment of Breast Cancer Risk Prediction Algorithms in Minority Populations. University of Wisconsin-Milwaukee. 2011-2012
- Collaborating Intelligent Health Information Systems Initiative (CIHISI). University of Wisconsin-Milwaukee, NSF-PFI Grant. 2007-2009
- Using Evidence-based Nursing Practices and EHR Decision Support to

Reduce Fall-related Patient Injuries in Acute Care. University of Wisconsin-Milwaukee and Aurora Health Care, AHRQ Grant. 2009
- Multilevel Comprehensive Models for the Analysis of the University Performance. Tehran University of Medical Sciences. 2003-2005
- Evaluating, Designing and Implementing a New Model for Health Care Delivery in the Trauma Center of Sina Teaching Hospital. Tehran University of Medical Sciences. 2004-2005
- An Ecological Study on the Relationship between Health Determinants and Health Status Indicators in Iran. Tehran University of Medical Sciences. 2005
- A Prevalence Study on Physical Diseases among Schizophrenic Patients Admitted to Major Psychiatric Hospitals in Tehran, Iran. Iran University of Medical Sciences. 1998
- Effects of Ramadan Fasting on Blood Biochemical Factors and Hypertension. Iran University of Medical Sciences. 1995

Experience in Developing Grant Proposals
- Predictive optimal anticlotting treatment for segmented patient populations (Peter Tonellato, PI), National Institutes of Health, 2013
- A Compute Cluster Supporting Multi-Disciplinary Genomics and Bioinformatics (Peter Tonellato, Michael Carvan, Sandra McLellan, Rebecca Klaper, Co-PIs), National Science Foundation MRI Program, 2013
- *In Silico* Pharmacogenetic Clinical Trial Design, Simulation, and Predicted Outcomes (Peter Tonellato, PI), National Science Foundation Smart Health and Wellbeing Program, 2012-2013
- Predicting clinical validity of bladder cancer nomograms (Peter Tonellato and Tracy Downs, Co-PIs), UW-Madison/UW-Milwaukee Intercampus Research Incentive Grants Program, 2012-2013
- Guiding Warfarin Clinical Trial Design Using Pharmacogenetic Simulations (Peter Tonellato, PI), University of Wisconsin-Milwaukee Research Growth Initiative, 2012-2014

Experience in Funded Projects
- Scientist on Predictive optimal anticlotting treatment for segmented patient populations (Peter Tonellato, PI), National Institutes of Health, 2013
- Scientist on Guiding Warfarin Clinical Trial Design Using Pharmacogenetic Simulations (Peter Tonellato, PI), University of Wisconsin-Milwaukee Research Growth Initiative, 2012-2014
- Scientist on Collaborative Intelligent Health Information Systems Initiative (CIHISI) (Sally Lundeen, PI), National Science Foundation, Partnerships for Innovation-PFI, 2007-2009.
- Scientist on Using Evidence-based Nursing Practices and EHR Decision Support to Reduce Fall-related Patient Injuries in Acute Care (Norma Lang, PI), AHRQ ACTION, 2009

Publications and Presentations
- **Ravvaz K**, Michalkiewicz M, Chi CL, Tonellato P. *Secondary use of electronic medical records to enhance in silico comparative effectiveness research: An application to anticoagulation health disparity*. JAMIA. *Preparation for re-submission*.
- Adams MB, Kaplan B, Sobko HJ, Kuziemsky C, **Ravvaz K**, Koppel R. *Learning from Colleagues about Healthcare IT Implementation and Optimization: Lessons from a Medical Informatics Listserv*. Journal of Medical Systems. 2015 Jan;39(1):157.

- Adams M, Kaplan B, **Ravvaz K**, Koppel R, Kuziemsky C. *What We See, Say, and Write: Qualitative Data Coding, Grounded Theory, and Other Analytic Approaches for Listservs, Records, and Patient Narratives*. <u>Tutorial</u> at American Medical Informatics Association Annual Conference. Nov. 2014. (*refereed*)
- **Ravvaz K**, Soulimi Y, Chi CL. *Pharmacogenetic Clinical Trial Simulation Workshop*. Center for Biomedical Informatics, Harvard Medical School. Apr. 2014. (*invited*)
- Adams M, Kaplan B, Koppel R, Kuziemsky C, **Ravvaz K**, Sobko H. *Insights from the Implementation Forum's Discussions: What Thirty Percent of AMIA Members Say about HIT Implementation and Use.* <u>Panel</u> at American Medical Informatics Association Annual Conference. Nov. 2013. (*refereed*)
- **Ravvaz K**, Tonellato P, Michalkiewicz M. *Development of an EMR-based database for anticoagulation/anticlotting outcome research.* Oral presentation at American Public Health Association Annual Meeting. Nov. 2013. (*refereed*)
- **Ravvaz K**, Michalkiewicz M, Tonellato P. *Secondary Use of a Statewide Electronic Health Record System for Patient-Centered Outcome Research: An Application to Anticoagulation Agents.* Poster presentation at American Medical Informatics Association Annual Conference. Nov. 2013. (*refereed*)
- **Ravvaz K**, Huang CY, Tonellato P, Patrick T. *In-silico analysis of the translation of warfarin pharmacogenetic knowledge*. Proceedings of 2012 College of Health Sciences Research Symposium, University of Wisconsin-Milwaukee, 2012:37. (*invited*)
- Miller T, **Ravvaz K**, Cimino JJ, Yu H. *An Investigation into the Feasibility of Spoken Clinical Question Answering*. Proceedings of the AMIA Symp 2011, American Medical Informatics Association; 2011:954-9. (*refereed*)
- Huang CY, Kos P, **Ravvaz K**, Tonellato P. *Meaningful Use of EHRs in Identifying Rural-Urban Health Disparities Using Pharmacogenetics-Based Clinical Avatar Simulations*. WPHA–WALHDAB Annual Conference, 2011. (*refereed*)
- Kos P, Huang CY, **Ravvaz K**, Tonellato P. *Venous thromboembolism and ischemic stroke risk disparities predicted in Milwaukee County*. WPHA-WALHDAB Annual Conference, 2011. (*refereed*)
- **Ravvaz K**, Patrick TB. *Measuring Semantic Density of Nursing Practice Recommendations*. Proceedings of the AMIA Symp 2009, American Medical Informatics Association, 2009:1014. (*refereed*)
- **Ravvaz K**, Senk P, Patrick T, Coenen A, Kim TY, Zhao H, Gaudioso C, Jensen K, Lang N. *Mapping Nursing Concepts to Ontologies for Evidence-Based Nursing*. Proceedings of the AMIA Symp 2008, American Medical Informatics Association, 2008:1105. ***This poster was nominated for an AMIA Distinguished Poster Award.*** (*refereed*)
- **Ravvaz K**, Patrick TB. *An Ethical Review of Euthanasia Web Sites*. Proceedings of the AMIA Symp 2007, American Medical Informatics Association, 2007:1088. (*refereed*)
- **Ravvaz K**, et al. *Tehran University of Medical Sciences and Health Services (English Prospectus)*. Tehran: TUMS Press, 2005.
- **Ravvaz K**. *Monthly International Relationships News Bulletin*, Tehran University of Medical Sciences, 2003-2005.
- **Ravvaz K**, Haghdoost AA, and Setayesh HR. *Effects of Ramadan fasting on blood biochemical factors and hypertension.* 9th Iran National Medical Student Scientific Congress, 1995. (*refereed*)

Databases
- Patrick T, **Ravvaz K**. Metadata Registry for "Risk for Fall Assessment" in USHIK (United States Health Information Knowledgebase): http://ushik.ahrq.gov/dr.ui.drOrg_View?System=mdr&OrgID=45&enableAsynchronousLoading=true
- **Ravvaz K**. WiAD: Wisconsin Anticoagulation/Anticlotting Database

Conference Participation and Attendances
- AMIA 2013 Annual Symposium
  American Medical Informatics Association, Washington DC, November 2013
- American Public Health Association Annual Conference, Boston, November 2013
- Genomics and Personalized Medicine Conference: Analysis and Clinical Implementation, Milwaukee, December 2012
- AMIA 2012 Annual Symposium
  American Medical Informatics Association, Chicago, November 2012
- Drug Development Collaborative Workshop, Milwaukee, May 2012
- College of Health Sciences Research Symposium
  University of Wisconsin-Milwaukee, May 2012
- Cancer Center Collaborative Workshop, Clinical and Translational Science Institute, Milwaukee, December 2011
- AMIA 2011 Annual Symposium
  American Medical Informatics Association, Washington DC, November 2011
- AMIA 2009 Annual Symposium
  American Medical Informatics Association, San Francisco, November 2009
- AMIA 2008 Annual Symposium
  American Medical Informatics Association, Washington DC, November 2008
- AMIA 2007 Annual Symposium
  American Medical Informatics Association, Chicago, November 2007
- AMIA 2006 Annual Symposium
  American Medical Informatics Association, Washington DC, November 2006
- Emergency Medicine Leadership Workshop
  Tehran University of Medical Sciences in collaboration with the George Washington University, Penn State University, Oregon Health and Science University, Johns Hopkins University and Loma Linda University, 2003

Honors and Awards
- Health Equity Leadership Institute Scholar, University of Wisconsin-Madison, 2014
- Biomedical and Health Informatics Research Institute (BHIRI) Award, University of Wisconsin-Milwaukee, 2013
- Health Equity Leadership Institute Scholar, University of Wisconsin-Madison, 2011
- University of Wisconsin-Milwaukee Travel Award, 2009, 2011
- University of Wisconsin-Milwaukee Chancellor's Award, 2006, 2007

Clinical Certification
- Medical Doctorate, Medical Council of Iran, 1998-present
- Registered Physician, Tehran, Iran, 1988-present

Societies
- Member, American Public Health Association (APHA)
- Member, American Medical Informatics Association (AMIA)
- Member, Global Health Informatics Working Group, American Medical Informatics Association (AMIA)

- Member, Wisconsin Public Health Association
- Member, Medical Council of Iran
- Co-Founder and President, Biomedical and Health Informatics Students Organization Association, University of Wisconsin-Milwaukee