Group and Dataset Comparisons for Optimization and Simulation-Based Approaches to Warfarin Treatment Protocols

Peter Ferm

Capstone Experience

BICB Program

University of Minnesota

Abstract

Warfarin is inexpensive and common, for being such, a powerful anticoagulant medication to treat and prevent blood clot disorders. Management of its important treatment measurements (INR and TTR) is constantly monitored due to the two-sided adverse effects. Narrow therapeutic windows coupled with high variability in dosage have been linked to clinical and genetic factors of an individual. The primary study under consideration utilized a simulation and optimization systematic approach and generated a study population of 1,478,930 simulated patients. These simulated patients were coined "clinical avatars" and are based on domain-knowledge, literature exploration, a Bayesian Model, and a Pharmacokinetic/Pharmacodynamic Model. A decision tree algorithm derived sub-populations to identify the factors that the clinical avatars could optimize by treatment protocols. A secondary analysis used a small random sample of 800 clinical avatars to investigate the effectiveness of warfarin treatment plans with respect to important clinical characteristics (age and BMI). Furthermore, how genetics and genomics are used in the Nursing environment. This capstone project aims to inspect and compare the varying sample sizes in the scope of the primary and secondary analyses. The Python language is used to produce visualizations and perform a logistic regression analysis to report and compare between groups and datasets.

Keywords: Warfarin; Atrial Fibrillation; Visualization; Logistic Regression; Bayesian Modeling; Effect Size; Odds Ratio; p-value Problems;

Introduction

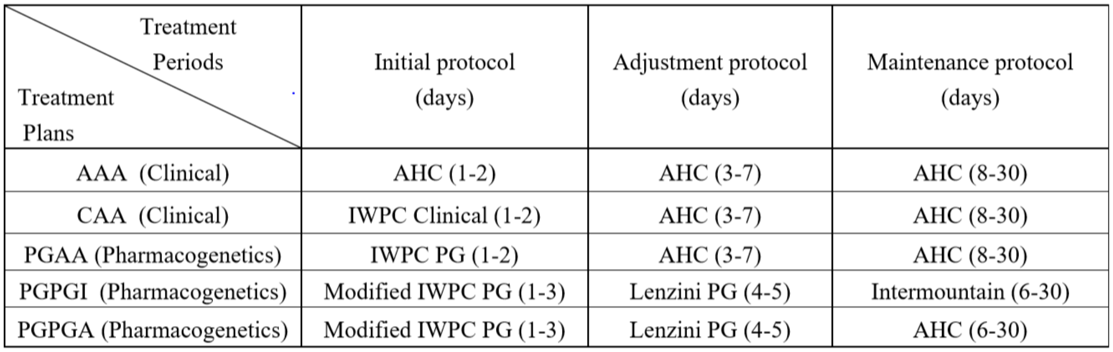
Warfarin—a frequently and highly efficacious used anticlotting agent—is a powerful oral medication to prevent and treat blood clotting disorders. It is widely used for those at risk of deep vein thrombosis, pulmonary embolism, and stroke.[5] Atrial fibrillation (AFib) is another common, however, a mild blood condition that warfarin is regularly prescribed for. AFib is a quivering or irregular heartbeat, where the atria do not contract correctly. Thus, the blood becomes static and causes it to pool. Consequently, this can lead to more serious heart complications.[5-8] Careful attention and persistence are mandatory when dealing with warfarin treatments. This is because of the narrow therapeutic windows combined with an individual's variability in sensitivity to warfarin. There are two-sided major adverse effects that are associated with warfarin treatments. Patients that are not within the therapeutic ranges are at risk of over-anticoagulation (i.e. causing bleeding), or under-anticoagulation (i.e. increased risk for strokes).[1-3, 6] To avoid these serious risk factors, two important measurements are monitored to ensure the safety and effectiveness for the patient. Blood tests are periodically drawn to assess the first metric, which is known as the International Normalized Ratio (INR). It is a standardized measurement for blood clots. The second measurement is the Time in Therapeutic Range (TTR). This quality measurement computes the percentage of time the INR measurement is in the therapeutic range. In other words, it is an indicator of the control and intensity of warfarin treatment. During warfarin treatments, the value at which an individual's INR should range is from 2.0 to 3.0, while normally not during warfarin treatment it is at 1.0. Well-controlled warfarin treatments are observed when TTR > 75%. When TTR < 60% it is indicative of poorly controlled warfarin treatments and when major bleeding events are more likely to occur. In general, the higher the TTR, the lower the two-sided risks.[1-3, 10, 11] The initiation of treatment is a crucial time for monitoring. There are over fifty protocol varieties for warfarin treatments. Currently, most healthcare providers adjust dosages depending on INR values. There are, nonetheless, treatment protocols that diversely combine the complexities of the patient's genetic, clinical, and demographic characteristics. There are mixed results whether the standardized clinical methods to warfarin protocols are optimal versus the pharmacogenomic (PG) methods. PG is the interactions of genetics with pharmacotherapy.[1-3, 26]

Although limited, clinical research has found that the high variability in INR response is connected to the patient's clinical and genetic characteristics. Information about an individual's clinical characteristics that heavily influence the sensitivity to warfarin, includes age and Body Mass Index (BMI). First, the older you are, the greater your risk of developing a condition such as AFib, especially individuals over the age of 65 years old.[7, 12] Second, the aging process can influence the body's pharmacodynamics (PD), which is the physiologic, molecular, and biochemical processes on how drugs affect the body, particularly when it comes to receptor binding. Third, the evolution of an individual's pharmacokinetics (PK), in other words, the movement of drugs into, throughout, and out of the body, changes at older ages.[31] A PK/PD model tells the relationship between dose and response.[20] Knowledge concerning a relationship between BMI and warfarin treatments have shown an association, where individuals' in the Obese category (BMI > 30) are more susceptible to over-anticoagulation.[13-15] Genetic characteristics of an individual can have a significant influence on the initial dosing of warfarin. Cytochrome P-450 2C9 (CYP2C9 alleles) and vitamin K epoxide reductase (VKORC1 alleles) are two enzymes identified to affect warfarin sensitivity. CYP2C9 gene provides instructions for making the principal enzyme in the metabolism of warfarin. Warfarin clinically exists with equal R and S enantiomers, and the CYP2C9 enzyme catalyzes metabolic clearances by way of oxidative reactions.[17] The VKORC1 gene encodes the vitamin K epoxide reductase enzyme. This enzyme essentially activates clotting proteins by converting one form of vitamin k into a different form of vitamin k.[18] There are three main genotype functional bins, that are included when classifying a patient's sensitivity to warfarin in the initial period of treatment. They are labeled as normal, sensitive, and highly sensitive responders. The CYP2C9/VKORC1 genetic testing can help healthcare providers predict these genotype functional bins for their patients. [5, 9]

When deciding warfarin treatment protocols many factors are involved for both the healthcare provider and the patient they are treating. Individual patients have a complex web of practical concerns and interdependent issues to untangle when making important healthcare and medical decisions. Some concerns include geography, communication, insurance coverage, and overall expenses.[19] Likewise, healthcare providers' and hospital administrators' complex decisions are based on improving treatment outcomes for a sundry collection of patients along with interoperability needs, Health Insurance Portability and Accountability Act (HIPAA) compliance, and budgetary constraints. Ideally, precision medicine could allow healthcare providers to treat the entirety of their patient population based on their specific clinical and genetic characteristics. Due to the numerous factors involved in healthcare treatment and management, it is not always practical or feasible for the patient population.

This capstone project is based on the research of a primary study[1-3] that took on these complex healthcare factors in the scope of warfarin treatment protocols. Dr. Chih-Lin Chi is a principal investigator, here at the University of Minnesota, on this collaborative study. The objective of the study was to find a cost-effective and realistic means for designing and evaluating precision-driven warfarin treatment protocols. The groundwork for their objective was cemented with the development of a comprehensive system of computational models that created a simulated warfarin clinical trial framework. To ensure the accuracy and cogency of the simulated framework, predictions of TTR were found to be directly analogous to TTR results of a real-world warfarin clinical trial.[26] The simulation was based on the published CoumaGen clinical trial. This was a randomized clinical trial that compared 101 patients undergoing a PG-guided protocol to 99 patients undergoing standardized clinical-guided protocol.[29] The primary study adapted their preliminary work as part of their methodology, to produce data for the primary study. The first step was to create a simulated patient population. The patient population that fit the scope of the study was the real-world patient data from the Aurora Health Care (AHC). The AHC serves communities throughout eastern Wisconsin and northern Illinois. Ten years (2002-2012), there were 14,206 AFib patients taking warfarin that were deemed suitable for the study. A core functionality of electronic medical records (EMRs) is storing clinical information and data. With Institutional Review Board approval, the AFib patient's EMRs were de-identified and the necessary clinical information was obtained. Next, the creation of the simulated patients, which were termed as 'clinical avatars', was based on computational approaches and a deep-dive into domain knowledge. The former, used the TETRAD[30] program, to execute a Bayesian model and stochastic models to match the patterns of the clinical information extracted from the EMRs. The latter did an expansive review of genetic tests, the CYP2C9 and VKORC1 genes, and the knowledge of their relationships with the clinical characteristics. The literature review was necessary due to the limited real-world genetic tests and limited storage of patient genotypes in their corresponding EMRs. After doing a statistical power analysis it was estimated the need to create 100 times the clinical avatars compared to the real-world AFib patients. Eventually, there were 1,478,930 clinical avatars in totality for the study population dataset. The simulated treatment values for each clinical avatar were predicted by developing a machine learning technique for a warfarin PK/PD model. Optimization approaches at an individual level helped determine the largest sub-populations that also had minimized two-sided adverse effects. A decision algorithm and a decision tree were built to identify the maximum count of clinical avatars that could benefit from the treatment protocol.[1-3]

There were 30-day and 90-day simulated periods, in which each clinical avatar went through five warfarin protocol treatment simulations (90-day period: 1,478,930 *clinical avatars* x 5 *protocols*; 30-period: 1,478,930 *clinical avatars* x 5 *protocols*). The overall study population dataset storing 7,394, 650 simulations processes for each simulation period, therefore totaling 14,789,300 data entries. These warfarin treatment protocols included two clinical-based protocols (AAA, CAA) and three PG protocols (PGAA, PGPGI, PGPGA).



Treatment Periods Treatment Plans	Initial protocol (days)	Adjustment protocol (days)	Maintenance protocol (days)
AAA (Clinical)	AHC (1-2)	AHC (3-7)	AHC (8-30)
CAA (Clinical)	IWPC Clinical (1-2)	AHC (3-7)	AHC (8-30)
PGAA (Pharmacogenetics)	IWPC PG (1-2)	AHC (3-7)	AHC (8-30)
PGPGI (Pharmacogenetics)	Modified IWPC PG (1-3)	Lenzini PG (4-5)	Intermountain (6-30)
PGPGA (Pharmacogenetics)	Modified IWPC PG (1-3)	Lenzini PG (4-5)	AHC (6-30)

Table 1: Five warfarin protocol treatment simulations and their respective clinical and/or genotypic based protocols and treatment periods.[2]

Each warfarin treatment plan whether clinical or PG, simulated through three treatment periods as seen in **Table 1**. The protocol coded 'AHC' refers to the current warfarin protocols practiced at Aurora Health Care. 'IWPC' is the current protocols from the International Warfarin Pharmacogenetics Consortium.[21] 'Intermountain' is the warfarin protocols based on the Intermountain Healthcare system. 'Lenzini' are protocols based on a previously published study into warfarin dosing.[22]

Miki Dahlin is an honor student at the University of Minnesota's School of Nursing. Miki with the advisement from Dr. Chih-Lin Chi performed a secondary analysis of the primary study.[4] The purpose of this secondary analysis was to interpret the impact of patient characteristics (age and BMI), on the effectiveness of clinical versus PG treatment protocols, measured by the TTR, for the 30-day treatment-simulation data. Moreover, the secondary analysis was to grasp how the field of nursing relates to genetics and genomics. The secondary analysis contains a random sample of 800 clinical avatars, from the primary study population analytical cohort. This random sample size was originally used for testing machine learning algorithms and was appropriate for the context of the secondary analysis. Similarly, this sample size went through each five distinct treatment protocols (800 *clinical avatars* x 5 *protocols*) to have a total of 4000 data entries. Descriptive statistics from the random sample investigated demographics for both clinical and genetic characteristics. There were three aims for evaluating the overall goal. A Chi-squared test of homogeneity was used to examine the three aims. The three aims are described in detail in the following sections.

The objective of the capstone project is to provide data validation and inspection to the secondary analysis and primary study through group comparisons for each dataset. Assessing data analysis quality is an essential task in any research project. The process of comparing various components of the analyses helps support the results and potentially find areas to improve on. Writing scripts that load in the data, wrangle the desired data, and implement computational and statistical methods can help towards a better understanding of these complex healthcare questions and thus maximizing overall healthcare outcomes.

Methods

For this capstone project, two datasets are used in multiple analyses. Therefore, to avoid confusion and clarify which dataset is being referenced, I have defined terms to represent the dataset under consideration. *Dataset 1* represents the population size of 1,478,930 clinical avatars. *Dataset 2* is the sample size of 800 clinical avatars. To further achieve the capstone's goal three tasks are proposed. These tasks include:

1. Compare the descriptive statistics of relevant attributes, between *Dataset 1* and *Dataset 2*.

2. Investigate the three aims from the secondary analysis for *Dataset 1*, using a logistic regression approach, then compare to the results of the secondary analysis three aims of *Dataset 2*.

    - The three aims from the secondary analysis included [4]

        1. To determine if there is a significant difference between treatment plans in the percentage of clinical avatars with a TTR ≥ 75%.

        2. To explore potential significant differences per treatment plan in the percentage of clinical avatars with a TTR ≥ 75% by age ≥ 65 years old, and age < 65 years old.

        3. To explore potential significant differences per treatment plan in the percentage of clinical avatars with a TTR ≥ 75% by weight category; BMI < 18.5 (Underweight), BMI ≥ 18.5 and < 25 (Normal), BMI ≥ 25, and <30 (Overweight) and BMI ≥ 30 (Obese).

3. Review the primary study's decision support algorithms to compare the *Dataset 1* decision tree to the *Dataset 2* decision tree.

The Python language works well in handling large raw datasets, and therefore it was fitting to address the objective and appropriate tasks. *Dataset 1* clinical avatars are stored in a pipe-delimited text file (~3.14 Gigabytes). *Dataset 1* stored thirty attributes about the clinical avatars. The attributes contained the avatar's clinical characteristics, genetic characteristics, primary identifications (e.g. Protocol, ID number, Period), and simulated warfarin treatment outcomes. *Dataset 2* contained a smaller list of attributes that were appropriate for the secondary analyses. *Dataset 2* avatars are stored in an excel file (~554 Kilobytes). A data dictionary is a useful way to describe the information and grasp a better understanding of the clinical avatars. So, before reading-in the data, I formed a data dictionary (see **Table 2**) of the relevant attributes that were used in the secondary analysis and corresponding capstone tasks.

Attribute	Description	Values*
“ID”	**Discrete variable for a clinical avatar’s identification number**	**Ranges from 1000001 to 2500000**
“GENDER”	**Binary variable indicating a Male or Female clinical avatar**	**“M” = Male “F” = Female**
“AGE”	**Continuous variable for current age (years) of the clinical avatars during the simulation**	**Ranges between 18-102**
“RACE”	**Nominal variable for the race of the clinical avatars**	**“White” “African-American” “Asian” “American Indian/Alaskan” “Native Hawaiian/Other Pacific Islander”**
“HEIGHT”	**Continuous variable for the height (inches) of clinical avatars**	**Ranges between 46.1-85.0**
“WEIGHT”	**Continuous variable for the weight (lbs.) of clinical avatars**	**Ranges between 71-490**
“SMOKER”	**Binary variable if the clinical avatar is a smoker**	**“Y” = Yes “N” = No**
“AMI”	**Binary variable if the clinical avatar is taking the medication Amiodarone or not**	**“Y” = Yes “N” = No**
“BMI”	**Continuous variable for Body Mass Index for each clinical avatar**	**Ranges between 8.7-91.3**
“ttrIn”	**Continuous variable for the percentage of TTR for each clinical avatar**	**Ranges between 0 - 100**
Rosendaal_ttrIn	**Continuous variable using the Rosendaal method for the percentage TTR for each clinical avatar**	**Ranges between 0 - 100**
“CYP2C9”	**Nominal variable for Cytochrome P-450 2C9 alleles (i.e. gene variations)**	**“*1/*1” “*1/*2” “*1/*3” “*2/*2” “*2/*3” “*3/*3”**
“VKORC1G”	**Nominal variable for Vitamin K epoxide reductase alleles (i.e. gene variations)**	**“A/A” “G/A” “G/G”**
“Protocol”	**Nominal variable for the five distinct treatment plans that each clinical avatar went through**	**“AAA” “CAA” “PGAA” “PGPGA” “PGPGI”**
“Period”	**Nominal variable for two simulation periods each clinical avatar went through**	**“90day” “30day”**

Table 2: Data dictionary of relevant clinical avatar attributes. *Values are based on *Dataset1*.

The input file containing *Dataset 1* stored both the 90-day and 30-day simulation periods. The first pre-processing task to perform was to wrangle only the 30-day simulations periods. The reason is to make a consistent representation of *Dataset 2*, which only contained 30-day simulation periods. To do this a regular expression package (regex; 2020.2.20) and its methods were used to match and capture the pattern of all the 30-day lines. The next pre-processing task was to import the captured 30-day lines into parallel lists. These lists are data structures in the Python language and can be used to wrangle desired avatar information and subsequently perform statistical and computational approaches. *Dataset 2* was imported into a data frame using the Pandas package (pandas; 1.0.1) to likewise prepare the data to be analyzed.

Task 1

The first task was to compare descriptive statistics between datasets. The processing of the task began with parsing the data into the correct data types and format. The descriptive statistics in the secondary analysis were based on counts and percentages of avatars attributes. Prior to changing the continuous variables into categories, the age and BMI attributes’ summary statistics were computed with the help of the frequently used NumPy package (NumPy; 1.18.1) and its methods. Next, the continuous attributes such as age, BMI, and ttrIn were looped into categories. The age category consisted of the age sub-groups '18-64' and '65+'. There were four Center for Disease Control (CDC) appropriate BMI category sub-groups, which were coded into, 'Underweight' (BMI < 18.5), 'Normal' (BMI ≥ 18.5 and < 25), 'Overweight' (BMI ≥ 25, and <30) and 'Obese' (BMI ≥ 30).[28] The ttrIn attribute was put into category sub-groups based on a conditional whether if they were in therapeutic range 'TTR ≥ 75%' or out of therapeutic range 'TTR < 75%'. To aid the comparison, visualization encompassing the avatar's relevant demographics, clinical characteristics, and genetic characteristics were created. Count plots visually help to inspect the distributions of counts and percentages for the larger *Dataset 1* compared to its smaller *Dataset 2*. This was an initial graphic to see if the *Dataset 2* was representative of *Dataset 1*.

Task 2

The second task's goal was to compute and compare inferential statistics of the secondary analysis three aims. The first step involved processing the protocol, age, BMI, and ttrIn attributes into the necessary data types and formats, as performed in Task 1. With the attributes in the correct format, it was time to wrangle the desired attributes into the context of aims 1, 2, and 3 of the secondary analysis. Like Task 1, to help explore these distributions for each dataset, count plots were produced in the context of each aim. The overall objective for each aim was to compare the category sub-groups to each other and to see if there are any significant differences amongst them. As mentioned, the secondary analysis used a Chi-square test of homogeneity to compare the category sub-groups and reported null hypothesis significance testing (NHST) p-values. The null hypothesis for the Chi-square test for homogeneity, states that the distribution of the categorical variables is the same for each sub-group. For *Dataset 1*, however, this statistical approach was not suitable for the size of the dataset. This is because smaller samples using NHST applied to very large samples will produce significant results even if there are little or no effects. The reason the p-values converge towards zero as the sample size increase is due to how the test statistics are computed concerning the sample size. Therefore, in this scope, a "p-value problem" arises with *Dataset 1*.[23, 27]

Keeping this in mind, a logistic regression approach was utilized for the sub-group comparisons to inspect if any association exists within the contexts of the secondary analysis aims. Reporting magnitudes and range of effects along with NHST p-values allows for a responsible result. Logistic regression is a machine learning technique for classification used in predicting binary outcomes. A binary outcome in general terms implies there are only two possible outcomes to a certain situation, which is usually represented as a '1' for success and '0' for failure. The logistic regression is derived from implementing the logit function:

$$Logit(\mathrm{p}) = log(\frac{\mathrm{p}}{1-\mathrm{p}})$$

which transforms the probability (p) of an outcome into the log odds. The odds and probabilities are different representations of the same information. This transformation into the log odds allows for values to range from -∞ to ∞, and therefore are not constrained to the odds ranges of 0 to ∞, or the probabilities ranges of 0 to 1. The general form of the logistic regression model is:

$$\mathrm{Log}\ odds\ of\ outcome = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Where there is *n*, sub-group predictor variables x_1 to x_n. The β's are the regression coefficients associated with the *n* sub-group predictor variables.[34]

The first step of the implementing logistic regression analysis in Python is coding the categories into the binary values. It is important to note, each attribute and its sub-groups are fit to separate logistic regression analyses. In other words, the age, BMI, and protocol attributes each goes through a logistic regression analysis. In this capstone perspective, the outcome variable we are trying to predict is the ttrIn attribute. So, if the clinical avatar is in therapeutic range (TTR ≥ 75%) it is coded as '1', and if it out of therapeutic range (TTR < 75%) it is coded as a '0'. The predictor variables are the age, BMI, and protocol attribute sub-groups. The next step is to dummy code the predictor variables. Dummy variables are a way to numerical distinguish each of the sub-groups. Typically, '1' represents the presence and '0' represents the absence. The number of dummy variables depends on the number of sub-groups it can assume. If there are *k* different sub-groups, then there will be *k*-1 dummy variables. This is done to avoid the dummy variable trap, which causes a multicollinearity problem. Having *k* dummy variables is not necessary and adds no new information. To examine the effect of each sub-group they are compared to a chosen baseline sub-group. It is important to note since the baseline sub-group is what the other sub-groups are being compared against its results are just that of the log odds rather than a log odds ratio. The statsmodels package (statsmodels; 0.11.0) and its logistic regression logit function was fit upon the transformed outcome variable and it the corresponding dummied predictor variable sub-groups. The results of the logistic regression are on a log scale and thus are converted back into odds perspective with the help of the NumPy natural exponentiate function.

Results

Task 1

Counts and percentages of the *Dataset 1* categorical attributes and *Dataset* 2 categorical attributes are visually presented in **Figure 1** and **Figure 2** respectively. At first glance, the majority of proportions compared between the datasets appear to reflect similar distributions. There is an additional Race category sub-group in *Dataset 1*. It is the 'Native Hawaiian/Other Pacific Islander' sub-group, in which there were only a total of 150 clinical avatars. Summary statistics were computed on the age and BMI continuous attributes. These were two attributes of focus for the secondary analysis. Summary statistics allow for a better understanding of these attributes' spread and central tendencies of the clinical avatars in each dataset. **Table 3** and **Table 4** are the comparisons between the datasets age attributes. *Dataset 1* represented in **Table 3** and *Dataset 2* represented in **Table 4**. The central tendencies amongst the datasets are comparable and the spread likewise is too. *Dataset 1* BMI attributes spread, and locations are in **Table 5**, while *Dataset 2* BMI attributes spread, and locations are in **Table 6**. Here it is seen that there are similar results across the summary statistics amongst the datasets, except in the maximum BMI values. *Dataset 1* has an extremely high BMI value of 91.3 lbs./in^2. *Dataset 2* has a high BMI value, however, there is a noteworthy difference seen in the range and maximum values between the datasets. Overall, the numerical summaries in the data exploration and visualization of the pertinent attributes are approximately parallel in each dataset. *Dataset 2* appears to be a good representation of *Dataset 1.*

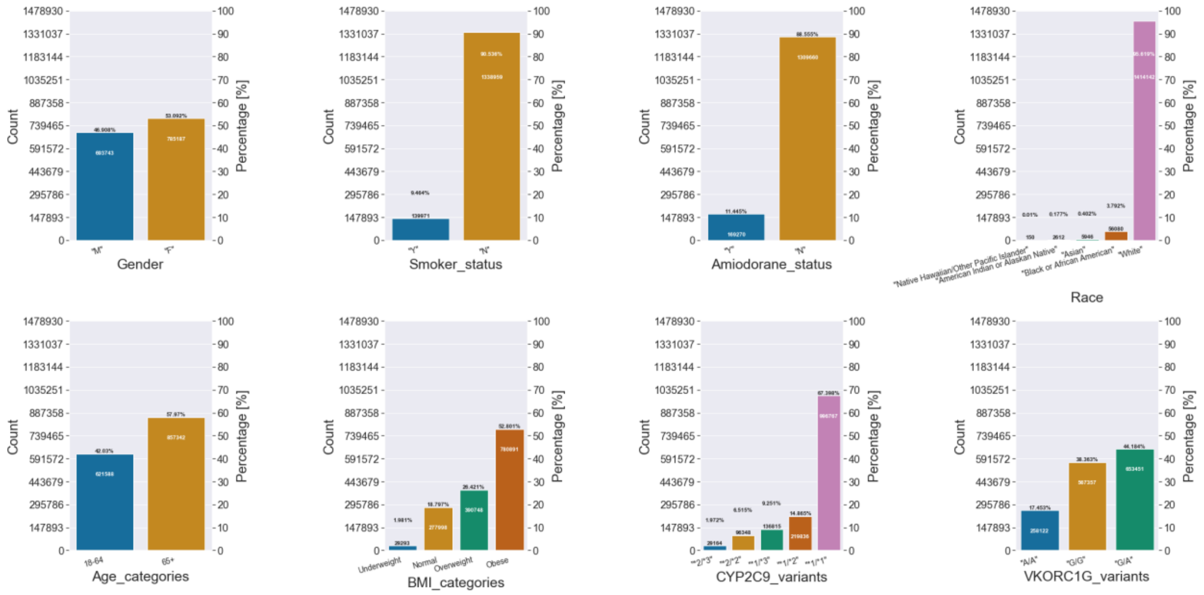


Figure 1: *Dataset 1* plots displaying counts and percentages of the clinical avatar's clinical, genetic, and demographic characteristics.

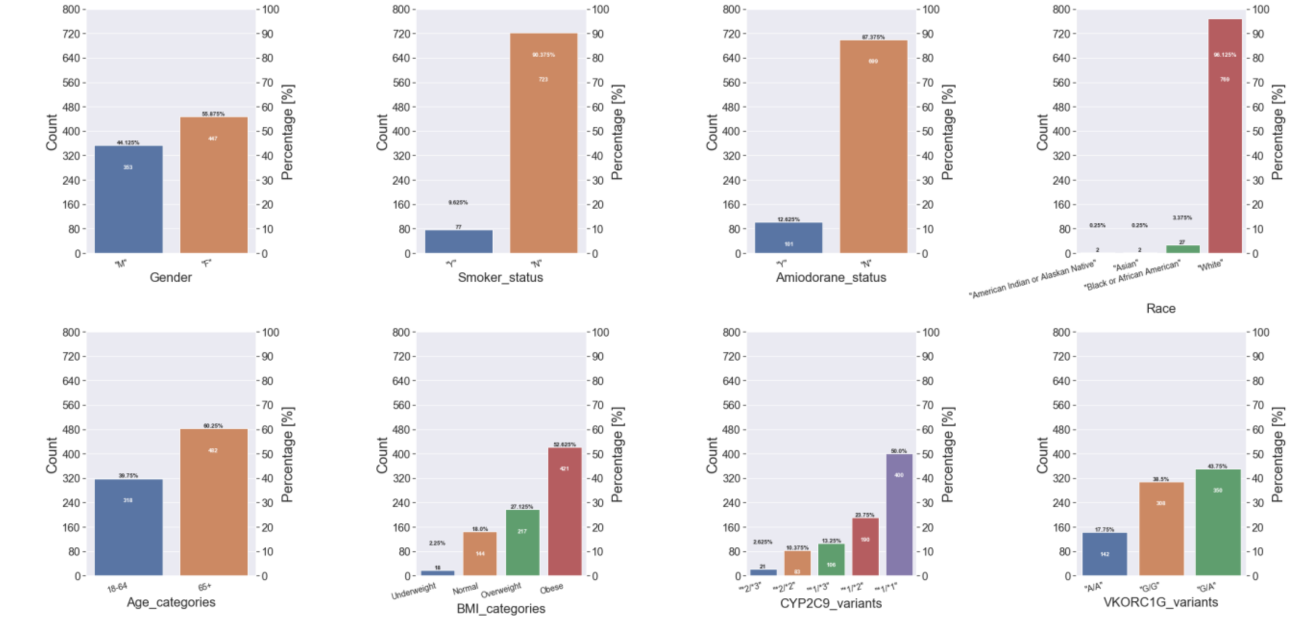
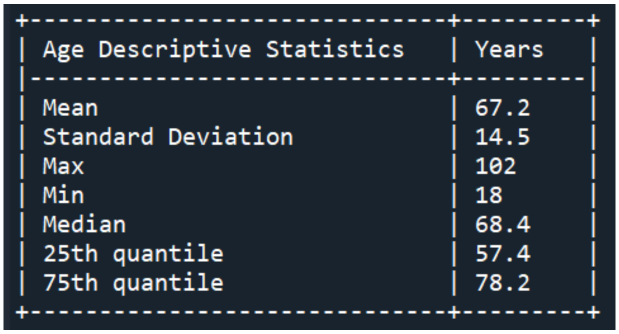
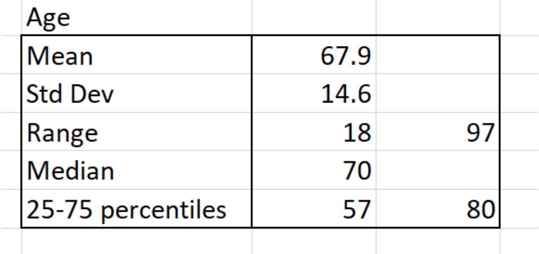


Figure 2: *Dataset 2* plots displaying counts and percentages of the clinical avatar's clinical, genetic, and demographic characteristics.



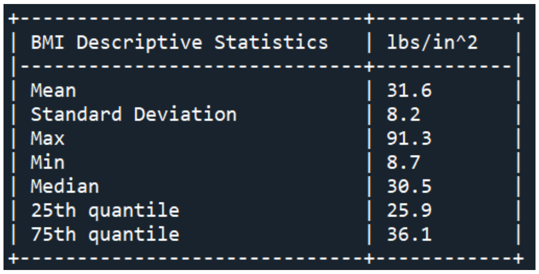
```
+------------------------------+---------+
| Age Descriptive Statistics   | Years   |
|------------------------------+---------|
| Mean                         | 67.2    |
| Standard Deviation           | 14.5    |
| Max                          | 102     |
| Min                          | 18      |
| Median                       | 68.4    |
| 25th quantile                | 57.4    |
| 75th quantile                | 78.2    |
+------------------------------+---------+
```

Table 3: Summary statistics of *Dataset 1* continuous age attribute.



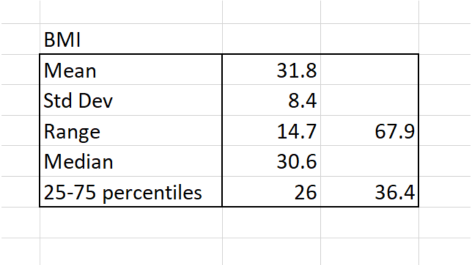
Age		
Mean	67.9	
Std Dev	14.6	
Range	18	97
Median	70	
25-75 percentiles	57	80

Table 4: Summary statistics of *Dataset 2* continuous age attribute.



```
+------------------------------+------------+
| BMI Descriptive Statistics   | lbs/in^2   |
|------------------------------+------------|
| Mean                         | 31.6       |
| Standard Deviation           | 8.2        |
| Max                          | 91.3       |
| Min                          | 8.7        |
| Median                       | 30.5       |
| 25th quantile                | 25.9       |
| 75th quantile                | 36.1       |
+------------------------------+------------+
```

Table 5: Summary statistics of *Dataset 1* continuous BMI attribute.



BMI		
Mean	31.8	
Std Dev	8.4	
Range	14.7	67.9
Median	30.6	
25-75 percentiles	26	36.4

Table 6: Summary statistics of *Dataset 2* continuous BMI attribute.

Task 2

The first step to executing Task 2 was to explore and visualize the age, BMI, and protocol category sub-groups being in or out of therapeutic range. Similar, to Task 1 this can give an initial inspection on how the different dataset sizes are represented, but now under the TTR quality measurement conditional. The protocol attribute categories are seen in **Figure 3** for *Dataset 1* and **Figure 4** for *Dataset 2*. The next visualization is of the age attribute categories, which are portrayed in **Figure 5** for *Dataset 1* and **Figure 6** for *Dataset 2*. The final visualizations for these therapeutic hues include **Figure 7** for *Dataset 1* and **Figure 8** for *Dataset 2*. The proportions compared amongst each dataset appear to follow the same distribution to each other.

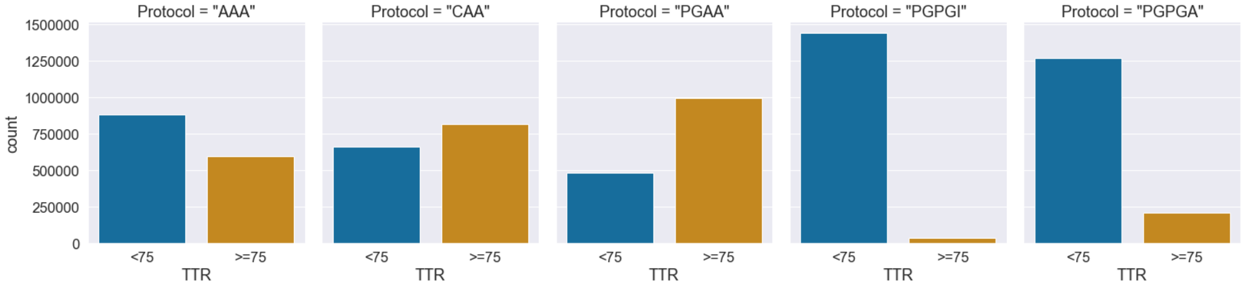


Figure 3: *Dataset 1* plots displaying counts of clinical avatars by the protocol in or out of therapeutic range.

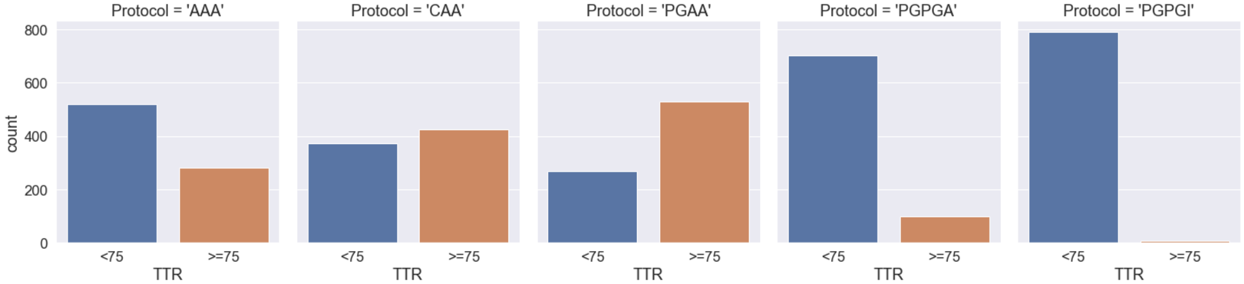


Figure 4: *Dataset 2* plots displaying counts of clinical avatars by the protocol in or out of the therapeutic range.

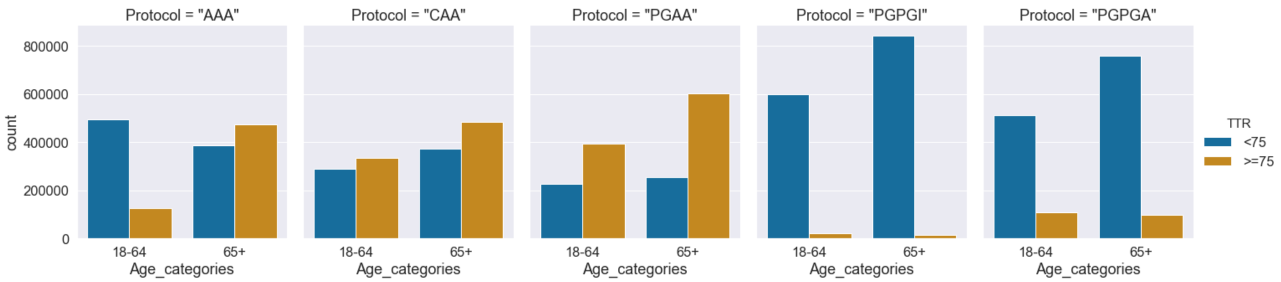


Figure 5: *Dataset 1* plots displaying counts of the clinical avatar's by age sub-groups in or out of therapeutic range for each treatment protocol.

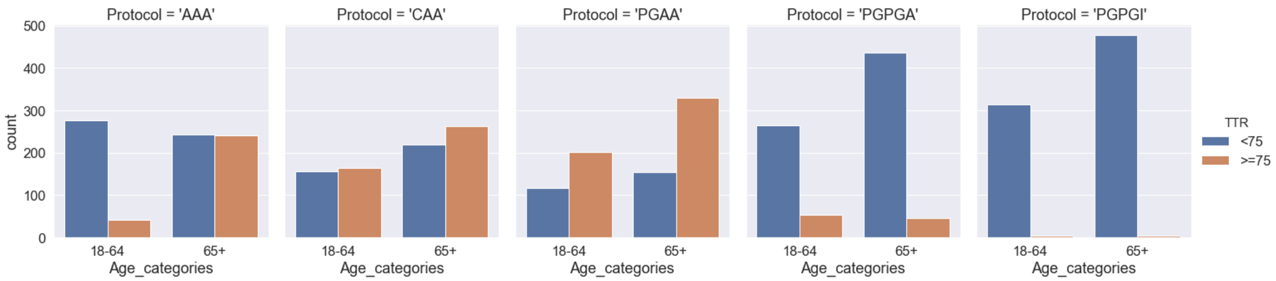


Figure 6 Dataset *2* plots displaying counts of clinical avatars by age sub-groups in or out of therapeutic range for each treatment protocol.

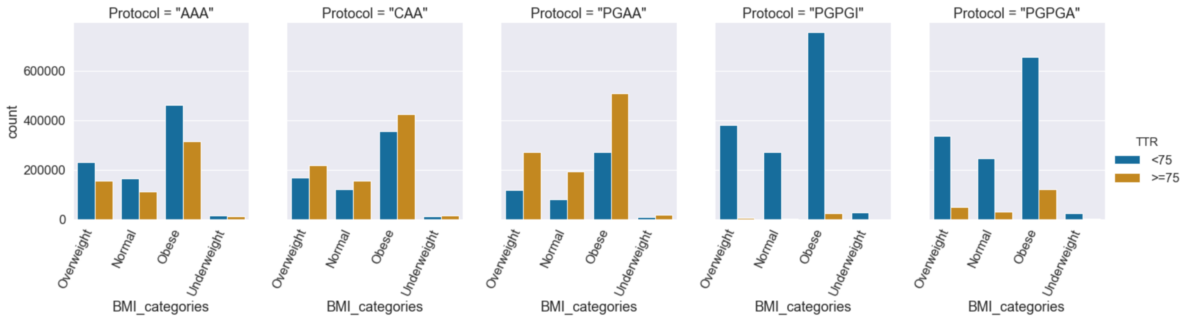


Figure 7: *Dataset 1* plots displaying counts and percentages by BMI sub-groups in or out of therapeutic range for each treatment protocol.

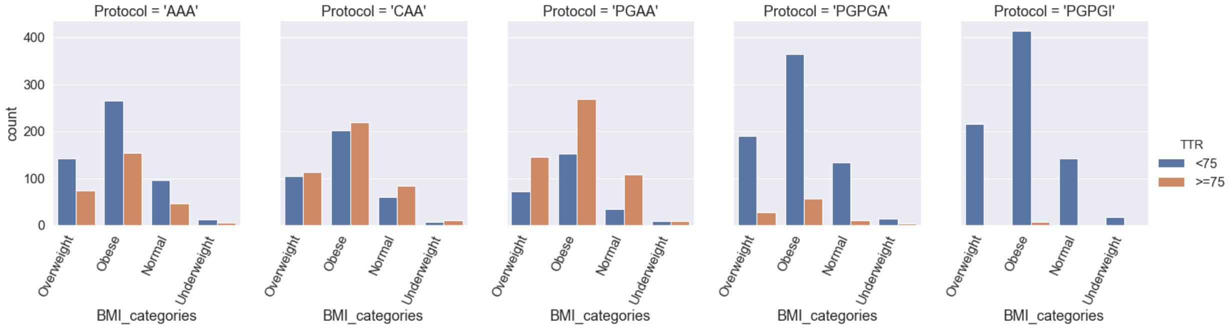


Figure 8: *Dataset 2* plots displaying counts and percentages by BMI sub-groups in or out of therapeutic range for each treatment protocol.

For the purpose of comparing inferential results for Task 2, it is necessary to give some background on examining the effect size. First, the effect size is a way to quantify the difference or associations between groups. The odds ratio (OR) is a way to determine effect size, and the corresponding 95% confidence interval (CI) quantifies the random error associated with our estimate. The CI is a measure of precision; with narrow CI being more precise and wider CI being less precise with more potential for error. The context of what you are measuring helps to determine how the effect size is estimated for the relative magnitudes. The estimations of the effect sizes are called indices. There are two main categories the indices fall into. These include looking at effect sizes between groups and looking at measures of association between attributes. When it comes to the OR, it is the latter, on how indices are estimated. Some common effect size magnitudes include when the OR = 1.5 (small), OR = 2 (medium) and OR = 3 (large).[24] Statistical significance, typically with the criteria of the p-value ≤ 0.05, with respect to the OR and its CI is tested against the null hypothesis; which states that the OR = 1. The CI estimate signifies if the OR is equal to 1 when the CI overlaps the value of 1 (e.g. if the CI is ranged from .99 to 1.01). This means the odds of the outcome are equally likely in each sub-group under consideration, and there is no association between the sub-group being tested to the baseline sub-group. When OR > 1, the sub-group being compared to the baseline is associated with higher odds of the outcome. Conversely, if OR < 1 the sub-group being compared to the baseline is associated with lower odds of the outcome.[32]

For *Dataset 1* a logistic regression analysis (see **Table 7**), produced the effect size estimates (i.e. OR and CI) and the odds for the protocol category's predictor sub-groups. These are based on their associations with the baseline sub-group, whose odds are also presented. This is compared to *Dataset 2* results using a Chi-square test of homogeneity (with the criterion of p-value ≤ 0.05) to compare if protocol sub-group percentages are the same amongst each other (see **Table 8**). The same iterative logistic processes are applied *Dataset 1* for age category sub-groups and the result are summarized in **Table 9**. Likewise, the same Chi-squared processes are applied to *Dataset 2* for age category sub-groups and the results are seen in **Table 10**. Lastly, the process is repeated for *Dataset 1* for each BMI category sub-group association in **Table 11**, compared to the results for *Dataset 2* BMI category sub-groups in **Table 12**.

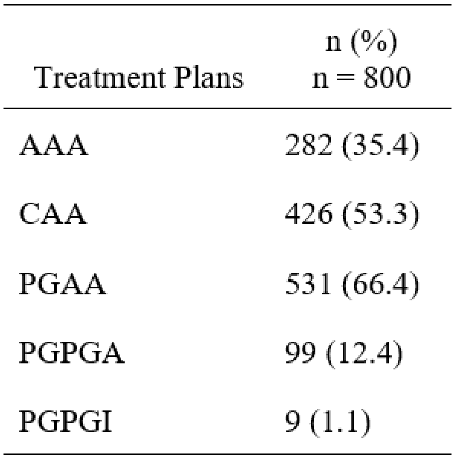
Recall **aim 1**, is to determine if there is a significant difference between treatment plans in the percentage of clinical avatars with a TTR ≥ 75%.

The first aim is focused on the protocol attribute. Results from **Table 7** for the logistic regression analysis, show all protocols, exhibited far below a small effect size when portending the odds of being in therapeutic range as compared to the baseline protocol. Although all the other protocols were well below the small magnitude size there were associations to the baseline protocol PGAA. They all were associated with lower odds when predicting the outcome of whether the clinical avatar's TTR is in therapeutic range as compared to the baseline. An example of an interpretation between the predictor sub-group and baseline sub-group is that clinical avatars in protocol AAA are 0.329 times less likely to be in therapeutic range than clinical avatars from the PGAA protocol. **Table 8** results saw a significant NHST p-value < 0.001 showing a difference amongst all the protocol sub-groups. Therefore, it is interpreted as there is a significant difference in the percentage of clinical avatars by protocol that are in therapeutic range. Here the PGAA protocol showing the highest percentage of clinical avatars in the therapeutic range.

When comparing between datasets, they appear to show the same pattern distribution of clinical avatars by protocol sub-groups being in therapeutic range. The PGAA protocol for both datasets and their corresponding analyses showed to be the most influential. It has the highest percentage in *Dataset 2* of clinical avatars in therapeutic range and greatest odds in predicting clinical avatars in therapeutic range for *Dataset 1*. The pattern follows suit in the order of CAA, AAA, PGPGA, to PGPGI, having the smallest effect sizes and odds and the lowest percentage of clinical avatars in therapeutic range in the respective datasets.

Protocol	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
AAA	0.329	(0.327, 0.330)	0.679
CAA	0.599	(0.597, 0.602)	1.24
*PGAA (baseline)	--	--	2.06
PGPGA	0.0789	(0.0785, 0.0794)	0.163
PGPGI	0.0122	(0.0121, 0.0123)	0.025

Table 7: Logistic regression outputs for the *Dataset 1* association between protocol treatment plans. *The baseline sub-group outputs just the odds of the clinical avatars being in therapeutic range, rather than an odds ratio.



Treatment Plans	n (%) n = 800
AAA	282 (35.4)
CAA	426 (53.3)
PGAA	531 (66.4)
PGPGA	99 (12.4)
PGPGI	9 (1.1)

Table 8: *Dataset 2* counts and percentages by protocol treatment plans in the therapeutic range. The NHST p-value < 0.001.[4]

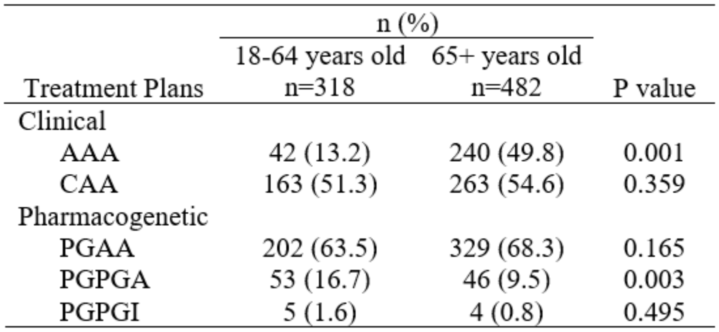
Recall **aim 2**, is to explore potential significant differences per treatment plan in the percentage of clinical avatars with a TTR ≥ 75% by age ≥ 65 years old, and age < 65 years old.

The second aim focused on the age attribute by protocol being in therapeutic range. **Table 9**, results show a small effect size and a medium effect size in two of the PG-guided protocols. The PGPGA protocol is associated with a small effect size where the OR = 1.65. This can be explained as clinical avatars in the age sub-group 18-64 years old undergoing the PGPGA simulated protocol are 1.65 times more likely to be in therapeutic range than the clinical avatars in the age sub-group of 65 plus years undergoing the PGPGA simulated protocol. Similarly, the PGPGI protocol had a medium effect size where the OR = 2.16. It is interpreted as clinical avatars in the age sub-group 18-64 years old undergoing the PGPGI simulated protocol are 2.16 times more likely to be in therapeutic range than the clinical avatars in the age sub-group of 65 plus years undergoing the PGPGI simulated protocol. All the other protocols and their corresponding age sub-groups showed to have effect sizes less than that of a small effect size magnitude. The AAA, CAA, and PGAA protocols all saw lower odds associated with 18-64 sub-group when compared to the 65+ sub-group. **Table 10** is showing a significant difference in the percentage of clinical avatars in therapeutic range for the age sub-groups in the AAA protocol (p-value = 0.001) and the PGPGA protocol (p-value = 0.003). There is a higher percentage of clinical avatars in therapeutic range in the 65+ sub-group in the AAA protocol, while there is a higher percentage of clinical avatars in therapeutic range in the 18-64 age sub-group for the PGPGA protocol.

Now between the datasets, it appears to comparable in the sense of having greater odds and higher percentages in the same age sub-groups. The AAA, CAA, PGAA, are all showing greater odds in the baseline 65+ sub-group for predicting clinical avatars in therapeutic range for *Dataset 1*. Furthermore, higher percentages in the 65+ sub-groups for *Dataset 2*. In the PGPGA and PGPGI protocols, we are seeing greater odds in the 18-64 sub-groups and higher percentages of 18-64 sub-groups. There seems to be some contest, regarding the effect sizes for *Dataset 1* compared to the significance for *Dataset 2*. In *Dataset 1* there is a medium effect size in the PGPGI protocol, whereas there is not a significant difference seen with the PGPGI protocol in *Dataset 2*. Rather *Dataset 2* is seeing a significant difference between the AAA protocol age sub-groups, while there is not a considerable effect size for the AAA protocol age sub-groups in *Dataset 1*.

AAA			
Age	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
18-64	0.206	(0.205, 0.208)	0.253
*65+ (baseline)	--	--	1.23
CAA			
Age	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
18-64	0.893	(0.887, 0.898)	1.16
*65+ (baseline)	--	--	1.30
PGAA			
Age	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
18-64	0.735	(0.730, 0.740)	1.73
*65+ (baseline)	--	--	2.36
PGPGA			
Age	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
18-64	1.65	(1.64,1.66)	0.213
*65+ (baseline)	-	-	0.129
PGPGI			
Age	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
18-64	2.16	(2.12, 2.21)	0.037
*65+ (baseline)	--	--	0.017

Table 9: Logistic regression outputs for *Dataset 1* association between age sub-groups. *The baseline sub-group output is just the odds of the clinical avatars being in therapeutic range, rather than an odds ratio.



	n (%)		
	18-64 years old	65+ years old	
Treatment Plans	n=318	n=482	P value
Clinical			
AAA	42 (13.2)	240 (49.8)	0.001
CAA	163 (51.3)	263 (54.6)	0.359
Pharmacogenetic			
PGAA	202 (63.5)	329 (68.3)	0.165
PGPGA	53 (16.7)	46 (9.5)	0.003
PGPGI	5 (1.6)	4 (0.8)	0.495

Table 10: *Dataset 2* counts and percentages by age sub-groups in the therapeutic range for protocol treatment plans; Accompanied by the Chi-square test's NHST p-values.[4]

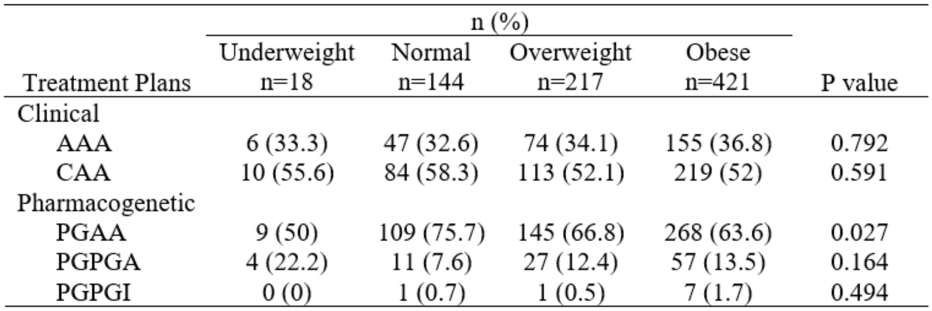
Recall **aim 3,** is to explore potential significant differences per treatment plan in the percentage of clinical avatars with a TTR ≥ 75% by BMI category; Underweight, Normal, Overweight, and Obese.

The third aim focused on BMI categories by protocol being in therapeutic range. The logistic regression outputs in **Table 11** see the Underweight and Normal sub-groups in the AAA protocol are showing no association with the Obese baseline sub-group. In other words, there are equal odds of predicting the clinical avatars being in therapeutic range in the Normal and Underweight sub-groups as there is for the Obese sub-group. Here we are seeing an estimate where the OR = 1 for both the Normal and Underweight sub-groups. The CAA and PGAA are close to having no associations of the sub-groups compared to the Obese baseline sub-groups, however, with the precise CIs, they are not overlapping the value of 1. Consequently, there are albeit, very slightly, higher odds when compared to the baseline Obese sub-group. The PGPGA and PGPGI protocols on the other hand are showing lower odds for all the sub-groups when compared to the Obese sub-group. The results in **Table 12**, show only the PGAA protocol having a significant difference in the percentage of clinical avatars in therapeutic range for its BMI sub-groups (p-value = 0.027).

The comparisons between datasets in the scope of BMI by protocols are not showing a similar pattern as in the two previous aims. This being the pattern where *Dataset 1* sub-groups order of high to low odds were matching the order of *Dataset 2* sub-groups high to low percentages. Looking at the comparison between datasets in a practical sense, there are some similarities. For instance, in the CAA protocol, the percentages amongst BMI sub-groups in *Dataset 2* are also are essentially the same. Likewise, for the CAA protocol, the odds in the BMI sub-groups are very close. In a statistical sense, however, the percentage of clinical avatars in *Dataset 2* is not statistically significant, though the *Dataset 1* results for this CAA BMI sub-groups technically are statistically significant by not having an OR = 1. This statistical significance difference between results in each dataset, potentially is due to how narrow the CIs are with the larger *Dataset 1,* while the smaller *Dataset 2* having a wider margin for error.

AAA			
BMI	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
Underweight	1.02	(0.993, 1.04)	0.692
Normal	0.998	(0.989, 1.01)	0.679
Overweight	0.990	(0.982, 0.998)	0.674
*Obese (baseline)	--	--	0.680
CAA			
BMI	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
Underweight	1.05	(1.02, 1.07)	1.25
Normal	1.09	(1.07, 1.10)	1.30
Overweight	1.09	(1.08, 1.10)	1.30
*Obese (baseline)	--	--	1.19
PGAA			
BMI	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
Underweight	1.17	(1.15, 1.21)	2.21
Normal	1.24	(1.23, 1.26)	2.34
Overweight	1.22	(1.21, 1.23)	2.29
*Obese (baseline)	--	--	1.88
PGPGA			
BMI	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
Underweight	0.524	(0.504, 0.550)	0.098
Normal	0.663	(0.654, 0.671)	0.124
Overweight	0.811	(0.802, 0.820)	0.151
*Obese (baseline)	-	-	0.187
PGPGI			
BMI	**Odds Ratio**	**95% Confidence Interval of Odds Ratio**	**Odds**
Underweight	0.303	(0.269, 0.341)	0.010
Normal	0.444	(0.430, 0.460)	0.014
Overweight	0.635	(0.620, 0.652)	0.020
*Obese (baseline)	--	--	0.032

Table 11: Logistic regression output for the *Dataset 1* association between BMI sub-groups. *The baseline sub-group output is just the odds of the clinical avatars being in therapeutic range, rather than an odds ratio.



	n (%)				
Treatment Plans	Underweight n=18	Normal n=144	Overweight n=217	Obese n=421	P value
Clinical					
AAA	6 (33.3)	47 (32.6)	74 (34.1)	155 (36.8)	0.792
CAA	10 (55.6)	84 (58.3)	113 (52.1)	219 (52)	0.591
Pharmacogenetic					
PGAA	9 (50)	109 (75.7)	145 (66.8)	268 (63.6)	0.027
PGPGA	4 (22.2)	11 (7.6)	27 (12.4)	57 (13.5)	0.164
PGPGI	0 (0)	1 (0.7)	1 (0.5)	7 (1.7)	0.494

Table 12: *Dataset 2* counts and percentages by BMI sub-groups in therapeutic range for protocol treatment plans; Accompanied with Chi-square test's NHST p-values.[4]

Task 3

A resulting outcome, from the primary study, was a decision tree algorithm, which supported decision rules to help individuals identify what sub-populations they belong with and in turn will help suggest a protocol treatment plan that minimizes the risk of the dangerous two-sided adverse effects. The decision tree algorithm is a clustering approach and was based on prior work of the prediction and optimization-based decision support system (PODSS) algorithm. The overall purpose of the PODSS algorithm is to promote actions that produce a higher probability of a desired outcome. These situations where we want to see the maximum outcome and the minimum amount of risks are seen across many domains. The PODSS algorithm captures important factors with a query input from a user and then outputs a suggested recommendation. It was important to remember the suggestion was not the be-all-end-all on how to make important decisions and resolving the trade-offs. The process of creating and using this decision tool started with designing a dataset to have independent and dependent variables to use as classifiers to build and train the predictive model. In other words, capturing the desired knowledge. Next is to extract the recommended information and optimize the maximum confidence level. Another important step to the recipe was building a validation map to help properly transform their data into probabilities.[19] In the scope of the primary study, the optimal level is based on the highest TTR across all five protocol treatment plans. This maximum TTR now becomes a label used in the decision tree algorithm. Two decision trees were derived from the decision support rules based on genetic and clinical characteristics. One tree was based on *Dataset 1*, and its results are represented in **Figure 9**. The other tree was mined from *Dataset 2* and **Figure 10** visualizes its results. The decision tree produced from *Dataset 1* eventually pruned the tree to have 12 decision support sub-populations, which are referred to as leaf nodes. *Dataset 2* produced 7 leaf nodes. PGAA is the most dominant protocol in *Dataset 1* with 37.8% of the proportion of prediction as to the optimal treatment plan. *Dataset 2* similarly saw PGAA as the dominant protocol garnering 43% of predictions. Both trees also saw the CAA protocol followed by the AAA protocol produce the next highest amount of nodes and proportion of predictions as the optimal treatment plans. Subsequently, the PGPGA protocol and then the PGPGI protocol round-up with the least amount of optimal predictions. These two protocols did not have any leaf nodes to equate as an optimal sub-population.[1,2]

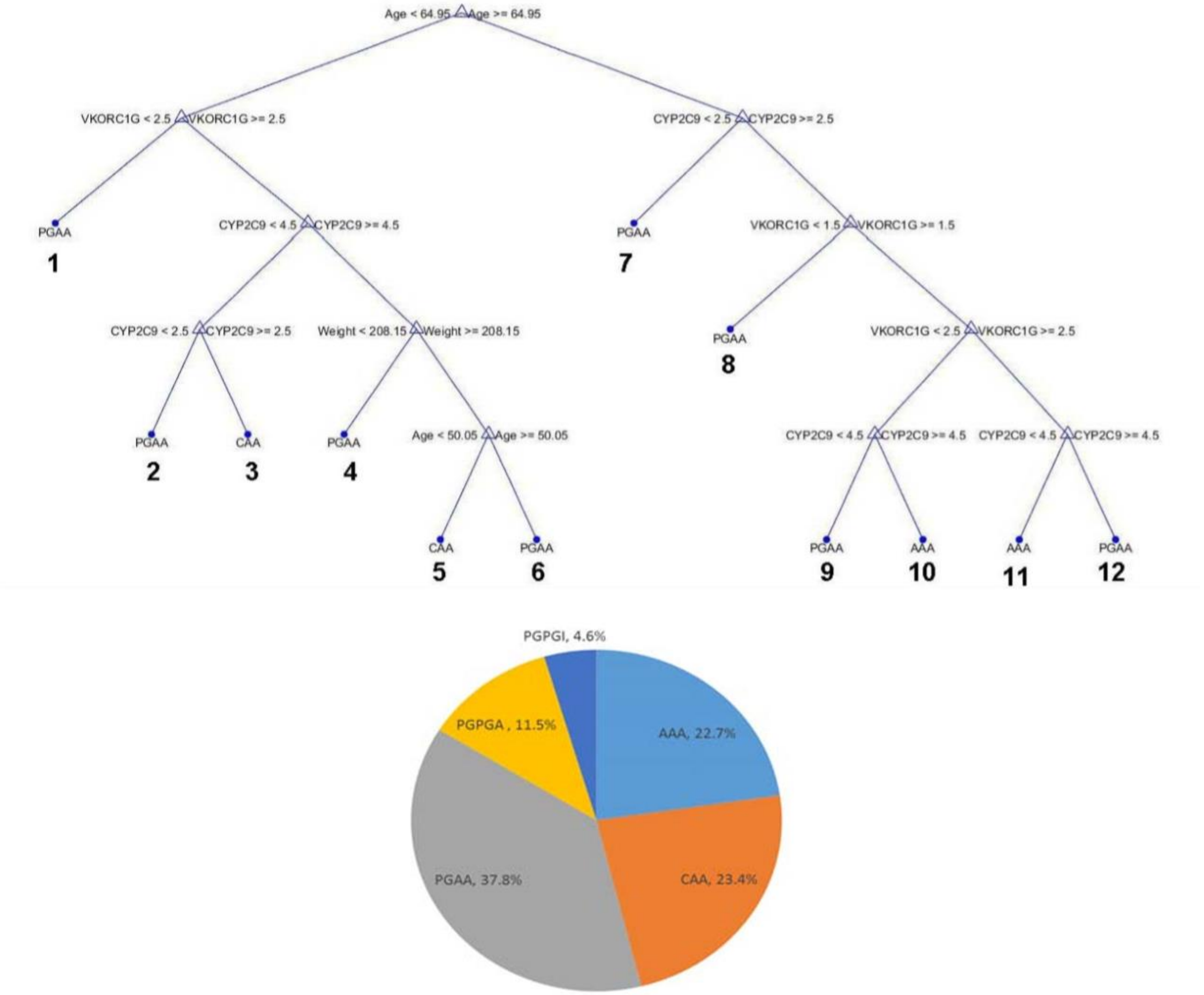


Figure 9: *Dataset 1* decision tree and pie-chart visualization of optimal sub-populations.[2]

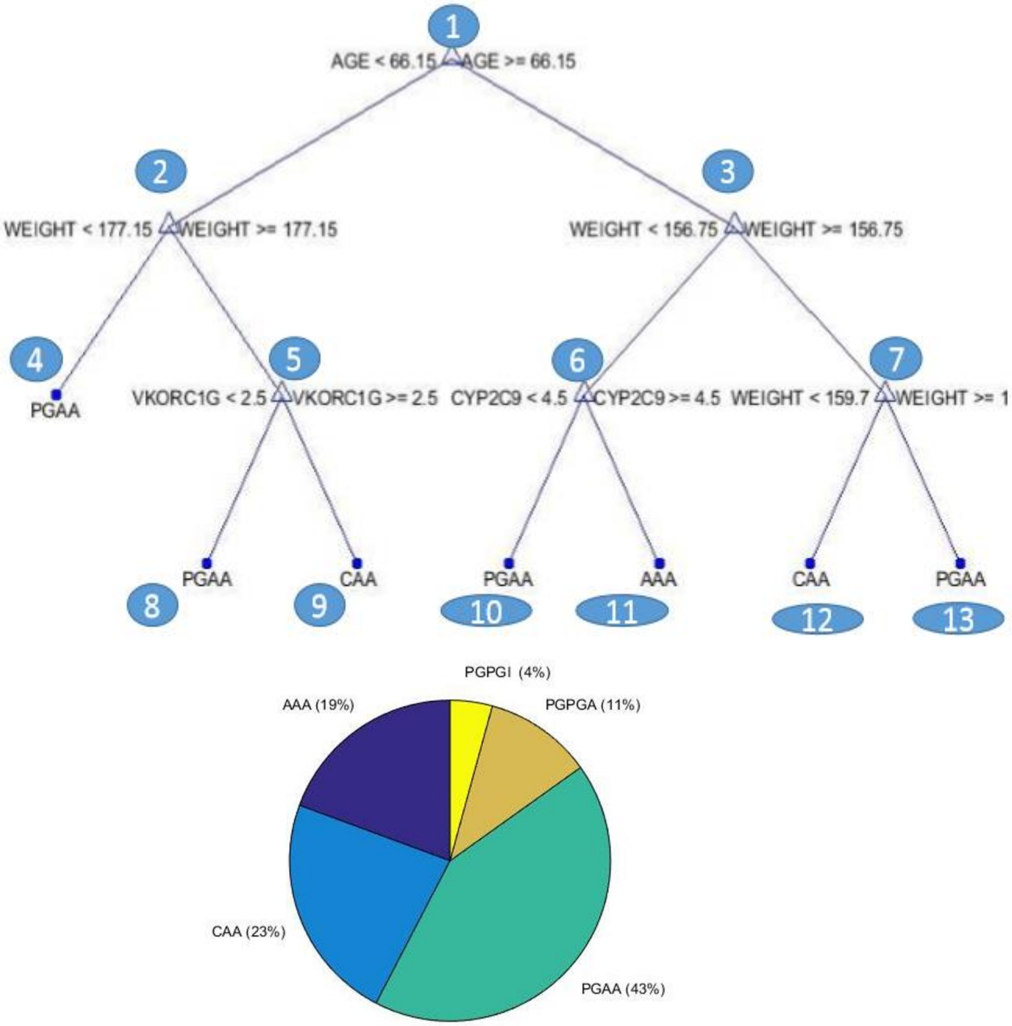


Figure 10: *Dataset 2* decision tree and pie-chart visualization of optimal sub-populations.[1]

Discussion

Discovering the warnings of the p-value problems with large samples when comparing between datasets in task 2 is amusingly like a Bayesian prior distribution. Bayesian methods use prior probabilities and their distributions to characterize our viewpoints about the world. Future consideration for this capstone project is applying a Bayesian approach to the sub-group comparisons in Task 2. Programming under a Bayesian lens could be beneficial due to the fact they do not fall victim to the p-value problems and increased precision estimates associated with large samples.[23] Using a larger sample size provides more precise estimates. CIs are computed from the same equation that generates p-values and is also affected by the sample size.[34] The narrower estimate may not detect a significant result when compared to smaller sample size with a wider estimate. In brief, Bayesian modeling can be summarized in three steps. First, with the help of data and assumptions about the origins of the data, a model is built using probability distributions. These distributions can range from knowing a lot about the data to knowing little. Next, Bayes' theorem is used to add data to our model, in order to see the logical results of adding both data and our assumptions. Finally, the model is put under scrutiny with domain-expertise and comparing several models to verify the model makes sense.[25]

Simulation frameworks and their techniques have a solid number of upsides to offer from the perspective of personalized anticoagulation therapies. A use-case scenario for a simulation approach, where the advantageous outcomes are apparent includes its cost-effectiveness of testing multiple treatment protocols, which are not always plausible in a practical setting. This plausibility can be due to a lack of genetic testing or sample size to work within a clinical environment. Clinical avatars can save time and money for embarking in a large traditional clinical trial. If a clinical trial is possible, simulation approaches can be the groundwork for designing a clinical trial and exploring optimal variables needed. As a decision-making reference, it potentially can reduce medical errors, by providing health care providers with more knowledge about the complexity of warfarin treatment and in turn decrease implementation of a daily complex clinical setting. Maintaining HIPAA standards is feasible when results from the simulation can be made publicly available. Many of these positive aspects of simulation-based approaches can help in the decision-making of a practical precision medicine choice. Some disadvantages of simulation approaches can be based on the models and what they are modeling. For example, warfarin has limitations that are difficult to model. These limitations may include interactions with other drugs or compliance with patients and physicians.[2, 26]

Overall, limitations in the primary study for modeling warfarin treatment protocols were based on the simplification of the models. Simplification in the PK/PD model potentially limited the ability to recognize the cofounders and covariates of warfarin protocols. Clustering approaches such as the one producing the decision trees, potentially do not present rules that intrigue healthcare providers. In the secondary analysis, the representation of the patient's characteristics to the greater population was a limiting factor. For, example the distribution of BMI for the avatars was not deemed to match the CDC general distribution of Wisconsin, one of the main locations, where the avatars are based from. The domain-knowledge concerning genetic variants for all the different races are not available or involved in the creation of the avatars. For the capstone experience, limitations occurred with the point estimates being affected by sample sizes and therefore causing biases, that made it difficult to compare results.

Conclusions

In this capstone experience, providing data inspection and comparisons to varying sample sizes, served as validation, and raised directions that need more exploration. Three tasks were proposed to help reach the objective of a better understanding of these complex warfarin treatment protocols and their respective managements. The first task involved comparing exploratory data analysis of varying sample sizes and validating the *Dataset 2* was an accurate representation of the *Dataset 1.* It appeared that this was the case and *Dataset 2* was a valid representation of *Dataset 1*. Some comparisons did not follow the same pattern. This was perhaps due to how each of the datasets was created. The second task involved predictions via inferential statistics. Whether or not influential attributes such as protocol, age by protocol, and BMI by protocol influenced the therapeutic outcome. Logistic regression helped evaluate the magnitude of this association for the categorical sub-groups. Here it was seen that the comparison between *Dataset 1* and *Dataset 2* showed both comparable and conflicting results in the attributes and category sub-groups under consideration. It appeared that the inconsistent results happened when the BMI category was involved in the comparison. The conflicting result was perhaps due to the difference in the sample size and how traditional statistics computations are influenced by sample size. A future consideration into Bayesian inferences would be a more efficacious approach for the second task. The third task involved comparing how varying sample sizes affected decision support rules and the resulting decision trees that were produced. There were related proportions and optimal sub-populations for the protocols seen for both decision trees for *Dataset 1* and *Dataset 2*.

This small capstone project experience was an opportunity where I gained valuable hands-on experience with informatics and computation. Learning new computational approaches and their benefits and limitations. It engages in a collaboration involving soft skills such as communication, networking, and project and time management. It provided active participation with data analysis, thus preparing myself for similar situations, which are necessary for the informatics and computational biology workplace today. It was important to learn how to adapt and adjust to one's current situations, to continually grow and learn. When wrangling data and doing data analysis it is important to continually check your progress. During the capstone experience, I found myself retracing my steps many times and finding new or different ways to solve the same task. Working in an iterative or non-linear fashion, allowed me to find silly coding mistakes or recognizing where bias was creeping into the results and methods. I accumulated more experiences with writing scripts, implementation of computational methods, data wrangling, interpretations, and report generation. Moreover, the capstone experience helps the continual growth in transforming data and information into potential knowledge.

Acknowledgments

Throughout the Capstone experience, several people have tremendously helped me along the way. I would like to start by thanking my advisor Dr. Chih-Lin Chi, for all the kindness, support, and experiences during all our meetings, whether remotely or in person. Our conversations have surely helped me grow and learn in a multitude of ways. My examining committee members, Dr. Yuk Sham and Dr. Noelle Noyes for joining and being a part of the capstone experience. I would like to thank them both for all the opportunities they have presented me throughout my graduate school career and all the graduate student advisement. I would like to thank, Miki Dahlin and Michelle Mathiason for engaging in collaboration and being very kind and resourceful. My fellow master's student Blake Jorgenson, for all the conversations (some with cold IPAs) and all his support throughout my graduate school career. My parents, for helping me in any way possible from back home in Montana. My soon to be wife Alice for everything she has done for me, especially putting up with me and getting after me to stay on track!

References

1. C.-L. Chi, K. Ravvaz, J. Weissert, and P. J. Tonellato, "Optimal decision support rules improve personalize warfarin treatment outcomes," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA, 2016, pp. 2594–2597, doi: 10.1109/EMBC.2016.7591261.
2. C.-L. Chi, L. He, K. Ravvaz, J. Weissert, and P. J. Tonellato, "Using simulation and optimization approach to improve outcome through warfarin precision treatment," in *Biocomputing 2018*, Kohala Coast, Hawaii, USA, 2018, pp. 412–423, doi: 10.1142/9789813235533_0038.
3. K. Ravvaz, J. A. Weissert, C. T. Ruff, C.-L. Chi, and P. J. Tonellato, "Personalized Anticoagulation: Optimizing Warfarin Management Using Genetics and Simulated Clinical Trials," *Circ Cardiovasc Genet*, vol. 10, no. 6, Dec. 2017, doi: 10.1161/CIRCGENETICS.117.001804.

4. M. Dahlin, “Examining the Efficacy of Pharmacogenetic and Clinical Warfarin Treatment Plans for Clinical Avatars in a Clinical Trial Simulation”, Thesis research project, University of Minnesota, School of Nursing, Minneapolis, MN 2019.

5. CYP2C9 & VKORC1/Warfarin Pharmacogenomic Lab Test - Center for Individualized Medicine - Mayo Clinic Research. Accessed May 19, 2020. https://www.mayo.edu/research/centers-programs/center-individualized-medicine/patient-care/pharmacogenomics/drug-gene-testing/cyp2c9-and-vkorc1-warfarin?_ga=2.58156522.114463769.1589681387-997540838.1589681387

6. Gomes T, Mamdani MM, Holbrook AM, Paterson JM, Juurlink DN. Persistence With Therapy Among Patients Treated With Warfarin for Atrial Fibrillation. *Arch Intern Med*. 2012;172(21):1687-1689. doi:10.1001/archinternmed.2012.4485

7. Atrial fibrillation - Symptoms and causes. Mayo Clinic. Accessed May 19, 2020. https://www.mayoclinic.org/diseases-conditions/atrial-fibrillation/symptoms-causes/syc-20350624

8. What is Atrial Fibrillation (AFib or AF)? www.heart.org. Accessed May 19, 2020. https://www.heart.org/en/health-topics/atrial-fibrillation/what-is-atrial-fibrillation-afib-or-af

9. Mega JL, Walker JR, Ruff CT, et al. Genetics and the clinical response to warfarin and edoxaban: findings from the randomised, double-blind ENGAGE AF-TIMI 48 trial. *The Lancet*. 2015;385(9984):2280-2287. doi:10.1016/S0140-6736(14)61994-2

10. Hasan SS, Sunter W, Ahmed N, et al. A comparison of warfarin monitoring service models. *Research in Social and Administrative Pharmacy*. 2019;15(10):1236-1242. doi:10.1016/j.sapharm.2018.10.029

11. Patel S, Singh R, Preuss CV, Patel N. *Warfarin*. StatPearls Publishing; 2020. Accessed May 19, 2020. https://www.ncbi.nlm.nih.gov/books/NBK470313/

12. Naccarelli, G. V., Varker, H., Lin, J., & Schulman, K. L. (2009). Increasing prevalence of atrial fibrillation and flutter in the United States. The American journal of cardiology, 104(11), 1534-1539.

13. Mueller, J. A., Patel, T., Halawa, A., Dumitrascu, A., & Dawson, N. L. (2014). Warfarin Dosing and Body Mass Index. Annals of Pharmacotherapy, 48(5), 584–588. doi: 10.1177/1060028013517541

14. Ogunsua, A. A., Touray, S., Lui, J. K., Ip, T., Escobar, J. V., & Gore, J. (2015). Body mass index predicts major bleeding risks in patients on warfarin. Journal of Thrombosis and Thrombolysis, 40(4), 494–498. doi: 10.1007/s11239-015-1226-2

15. Wallace, J. L., Reaves, A. B., Tolley, E. A., Oliphant, C. S., Hutchison, L., Alabdan, N. A., … & Self, T. H. (2012). Comparison of initial warfarin response in obese patients versus non-obese patients. Journal of Thrombosis and Thrombolysis, 36(1), 96–101. doi: 10.1007/s11239-012-0811-x

16. Higashi MK, Veenstra DL, Kondo LM, et al. Association Between CYP2C9 Genetic Variants and Anticoagulation-Related Outcomes During Warfarin Therapy. *JAMA*. 2002;287(13):1690-1698. doi:10.1001/jama.287.13.1690

17. Kim S-Y, Kang J-Y, Hartman JH, et al. Metabolism of R- and S-Warfarin by CYP2C19 into Four Hydroxywarfarins. *Drug Metab Lett.* 2012;6(3):157-164.

18. Reference GH. VKORC1 gene. Genetics Home Reference. Accessed May 19, 2020. https://ghr.nlm.nih.gov/gene/VKORC1

19. Chi C-L, Street WN, Ward MM. Building a hospital referral expert system with a Prediction and Optimization-Based Decision Support System algorithm. *Journal of Biomedical Informatics*. 2008;41(2):371-386. doi:10.1016/j.jbi.2007.10.002

20. Overview of Pharmacodynamics - Clinical Pharmacology. Merck Manuals Professional Edition. Accessed May 19, 2020. https://www.merckmanuals.com/professional/clinical-pharmacology/pharmacodynamics/overview-of-pharmacodynamics

21. International Warfarin Pharmacogenetics Consortium. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8), 753-764.

22. Lenzini P, Wadelius M, Kimmel S, et al. Integration of genetic, clinical, and INR data to refine warfarin dosing. *Clin Pharmacol Ther*. 2010;87(5):572-578. doi:10.1038/clpt.2010.13

23. Szucs D, Ioannidis JPA. When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Front Hum Neurosci.* 2017;11. doi:10.3389/fnhum.2017.00390

24. Sullivan GM, Feinn R. Using Effect Size—or Why the P Value Is Not Enough. *J Grad Med Educ*. 2012;4(3):279-282. doi:10.4300/JGME-D-12-00156.1

25. Martin O. *Bayesian Analysis with Python: Introduction to Statistical Modeling and Probabilistic Programming Using PyMC3 and ArviZ, 2nd Edition*. Packt Publishing Ltd; 2018.

26. Fusaro VA, Patil P, Chi C-L, Contant CF, Tonellato PJ. A systems approach to designing effective clinical trials using simulations. Circulation. 2013;127(4):517-526. doi:10.1161/CIRCULATIONAHA.112.123034. 9

27. Lin M, Lucas Jr. HC, Shmueli G, Too Big to Fail: Large Samples and the p-Value Problem. Information Systems Research. 2013; Articles in Advance, pp. 1 -12 :https://pdfs.semanticscholar.org/262b/854628d8e2b073816935d82b5095e1703977.pdf/

28. Center for Disease Control. (2015). Body Mass Index (BMI). Retrieved from https://www.cdc.gov/healthyweight/assessing/bmi/index.html

29. Anderson Jeffrey L., Horne Benjamin D., Stevens Scott M., et al. A Randomized and Clinical Effectiveness Trial Comparing Two Pharmacogenetic Algorithms and Standard Care for Individualizing Warfarin Dosing (CoumaGen-II). *Circulation.* 2012;125(16):1997-2005. doi:10.1161/CIRCULATIONAHA.111.070920

30. Scheines R, Spirtes P, Glymour C, Meek C, Richardson T. The TETRAD Project: Constraint Based Aids to Causal Model Specification. Multivariate Behav Res. 1998;33(1):65-117. doi:10.1207/s15327906mbr3301_3

31. Overview of Pharmacokinetics - Clinical Pharmacology. Merck Manuals Professional Edition. Accessed May 27, 2020. https://www.merckmanuals.com/professional/clinical-pharmacology/pharmacokinetics/overview-of-pharmacokinetics

32. Szumilas M. Explaining Odds Ratios. *J Can Acad Child Adolesc Psychiatry*. 2010;19(3):227-229.

33. Confidence Intervals and p-Values. Accessed May 28, 2020. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_RandomError/EP713_RandomError6.html#headingtaglink_3

34. Kirkwood, Betty R, Jonathan A. C. Sterne, and Betty R. Kirkwood. Essential Medical Statistics. Malden, Mass: Blackwell Science, 2003.