DISSERTATION

MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE APPROACHES TO THE ANALYSIS OF PHYSICAL ACTIVITY FROM WEARABLES AND BIOSENSORS IN CLINICAL TRIALS: APPLICATIONS OF CLUSTERING AND PREDICTION OF CLINICAL OUTCOMES

Submitted by
Vanja M. Vlajnic
Department of Systems Engineering

In partial fulfillment of the requirements
For the Degree of Doctor of Philosophy
Colorado State University
Fort Collins, Colorado
Summer 2022

Doctoral Committee:

Advisor: Steve Simske

Erika Miller
James Cale
Bradley Reisfeld

ABSTRACT

MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE APPROACHES TO THE ANALYSIS OF PHYSICAL ACTIVITY FROM WEARABLES AND BIOSENSORS IN CLINICAL TRIALS: APPLICATIONS OF CLUSTERING AND PREDICTION OF CLINICAL OUTCOMES

As human demographics continue to trend toward elderly, especially in advanced economies, the treatment of illness becomes more salient. Across many therapeutic areas, researchers examine potential treatments while incorporating novel technologies in an effort to prolong the years in which quality of life is achieved for patients around the world. In the area of cardiovascular disease, wearable and biosensor data is becoming increasingly used in order to compliment data traditionally collected from clinical trials. This work discusses a series of analytical approaches for the analysis of data from recent clinical trials in which accelerometry data from wearable devices were analyzed using clustering approaches (K-means and consensus clustering) and survival analyses (Cox proportional hazards and random survival forest) for the purposes of clustering patients and assessing their baseline clinical characteristics as well as for the prediction of clinical outcomes. Unique clinical phenotypes were identified within the patient aggregations as part of the clustering analyses. Furthermore, models were created with improved predictive accuracy for clinical outcomes of interest in the heart failure space. Taken collectively, the results from these analyses and the analytical approaches therein can be used to assess whether heterogeneous clinical subgroups of patients exist as well as further guide the clinical development programs.

ACKNOWLEDGEMENTS

I would like to thank all of my colleagues at CSU and in particular the Systems Engineering Department. None of this would be possible without my advisor, Steve Simske, and his guidance and mentorship over the years. I would also like to thank my committee members, Erika Miller, James Cale, and Bradley Reisfeld for all of their support.

Furthermore, I would like to thank my many Bayer colleagues who have been supportive of my goal to pursue a Ph.D. In particular, I would like to thank my previous manager, Dan Haverstock, for supporting my proposal to start this program and my current manager, Keith Bangerter, for fully supporting my work as well. I'd also like to thank my other Bayer colleagues Hermann Kulmann, Olga Marchenko, Torsten Westermeier, and Benedikt Egersdoerfer for the opportunities they have offered me over the years. I recognize I've been very lucky to have such great managers and mentors.

I would like to thank my friends and family for being understanding of the the late nights and the limited free time over the last few years. In particular, I would like to thank my wife, Jaclyn, for all the support she has provided me with. Her continued belief in me has strengthened my own belief in myself. Most of all I would like to thank God who has granted me with blessings and opportunities for success.

I recognize this work would not have been possible without the support, guidance, and mentorship of those above and many others. Thank you all. I hope I can leave a positive impact on those around me as you all have on me.

DEDICATION

I would like to dedicate this dissertation to my parents, Aleksandra and Miodrag Vlajnic, for all of the sacrifices they made in their own lives to help provide me with opportunities to make my dreams a reality.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iii
DEDICATION iv
LIST OF TABLES vii
LIST OF FIGURES viii

Chapter 1 Introduction 1
1.1 Clinical trials 1
1.2 Cardiovascular disease and heart failure 3

Chapter 2 Datasets 8
2.1 Background on dataset 8
2.2 Physical activity data from wearables and biosensors 8
2.3 Physical activity duration and intensity across the trial 9
2.4 Medical history and baseline characteristic variables 10
2.5 Clinical outcomes 12

Chapter 3 Methods: K-Means Clustering 14
3.1 Aggregation of patients based on physical activity duration and intensity through K-means clustering 14
3.2 Methods 14
3.3 Results 15
3.4 Discussion and interpretation 18

Chapter 4 Methods: Consensus Clustering 20
4.1 Aggregation of patients based on physical activity duration and intensity through Consensus clustering 20
4.2 Methods 20
4.3 Results 21
4.4 Discussion: Limitations of clustering approach 26

Chapter 5 Methods: Cox Proportional Hazards Model 27
5.1 Introduction to predictive models of clinical outcomes 27
5.2 Cox Proportional Hazards Model and Method 27
5.3 Results for time to first heart failure hospitalization based on Cox PH model 28
5.4 Results for the composite of time to first heart failure hospitalization or urgent heart failure visit based on Cox PH model 33
5.5 Results for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model . . 38
5.6 Discussion based on Cox PH model 43

Chapter 6 Methods: Random Survival Forest 45

6.1 Results for time to first heart failure hospitalization based on Random Survival Forest 48
6.2 Results for the composite of time to first heart failure hospitalization or urgent heart failure visit based on Random Survival Forest 58
6.3 Results for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Random Survival Forest 68
6.4 Discussion for Random Survival Forest 80

Chapter 7 Discussion and clinical interpretation of findings 82
7.1 Limitations and future development 84

Bibliography 87

Appendix A Programming Code 95

LIST OF TABLES

3.1 K-means clustering output with respect to baseline medical history and demographics 17

4.1 Descriptive statistics of activity intensity and duration across clusters identified by consensus clustering 23
4.2 Baseline and demographic data across clusters identified by consensus clustering . . . 24

5.1 Cox Proportional Hazards model output for time to first heart failure hospitalization . . 29
5.2 Cox Proportional Hazards model with backwards selection output for time to first heart failure hospitalization 30
5.3 Cox Proportional Hazards model final reduced output for time to first heart failure hospitalization 31
5.4 Cox Proportional Hazards model output for the composite of time to first heart failure hospitalization or urgent heart failure visit 34
5.5 Cox Proportional Hazards model with backwards selection output for the composite of time to first heart failure hospitalization or urgent heart failure visit 35
5.6 Cox Proportional Hazards model final reduced output for the composite of time to first heart failure hospitalization or urgent heart failure visit 36
5.7 Cox Proportional Hazards model output for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model 39
5.8 Cox Proportional Hazards model output for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model 40
5.9 Cox Proportional Hazards model final reduced output for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model 41

6.1 Random survival forest minimal depth and VIMP scores for clinical outcome of heart failure 50
6.2 Random survival forest minimal depth and VIMP scores for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit . . . 60
6.3 Random survival forest minimal depth and VIMP scores for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death 70

LIST OF FIGURES

3.1 Plot visualizing optimal number of clusters based on the grid search optimization . . . 15
3.2 K-means patient cluster plot visualizing the distribution of patients into the cluster groupings 16

4.1 RCSI plot visualizing the RCSI across K clusters 22
4.2 Entropy plot visualizing the information entropy graph 22

5.1 Schoenfeld residuals for Cox Proportional hazards model for time to first heart failure hospitalization 32
5.2 Schoenfeld residuals for Cox Proportional hazards model for the composite of time to first heart failure hospitalization or urgent heart failure visit 37
5.3 Schoenfeld residuals for Cox Proportional hazards model for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model 42

6.1 Random survival forest VIMP output for clinical outcome of heart failure hospitalization 49
6.2 Random survival forest minimal depth output for clinical outcome of heart failure hospitalization 51
6.3 Random survival forest minimal depth and VIMP output for clinical outcome of heart failure hospitalization 52
6.4 Random survival forest final reduced VIMP output for clinical outcome of heart failure hospitalization 53
6.5 Partial dependence plot for baseline NT-proBNP for clinical outcome of heart failure hospitalization 54
6.6 Partial dependence plot for baseline NYHA class for clinical outcome of heart failure hospitalization 55
6.7 Partial dependence plot for prior history of heart failure hospitalization class for clinical outcome of heart failure hospitalization 56
6.8 Partial dependence plot for activity duration for clinical outcome of heart failure hospitalization 57
6.9 Partial dependence plot for activity intensity for clinical outcome of heart failure hospitalization 58
6.10 Random survival forest VIMP output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit 59
6.11 Random survival forest minimal depth output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit 61
6.12 Random survival forest minimal depth and VIMP output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit . . . 62
6.13 Random survival forest intermediate reduced VIMP output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit . . . 63
6.14 Random survival forest final reduced VIMP output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit 64

6.15 Partial dependence plot for baseline NT-proBNP for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit 65
6.16 Partial dependence plot for baseline NYHA class for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit 66
6.17 Partial dependence plot for baseline activity duration for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit 67
6.18 Partial dependence plot for baseline activity intensity for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit 68
6.19 Random survival forest VIMP output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death . . . 69
6.20 Random survival forest minimal depth output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death 71
6.21 Random survival forest minimal depth and VIMP output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death 72
6.22 Random survival forest intermediate reduced VIMP output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death 73
6.23 Random survival forest final reduced VIMP output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death 74
6.24 Partial dependence plot for baseline NT-proBNP for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death 76
6.25 Partial dependence plot for baseline NYHA class for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death 77
6.26 Partial dependence plot for baseline activity duration for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death 78
6.27 Partial dependence plot for baseline activity intensity for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death 79

Chapter 1

Introduction

1.1 Clinical trials

The science of clinical trials has continued to expand immensely since Dr. James Lind ran the first modern-day clinical trial in 1747, almost a full 275 years ago [1]. In what became known as the first contemporary clinical trial, Dr. Lind was interested in examining the possible treatment options for 12 sailors suffering from scurvy on board of his ship at sea. He randomly assigned each set of 2 an additional supplementary food as part of their controlled diet to assess if there were any changes in the clinical outcome of scurvy. Dr. Lind had perhaps arguably setup the world's first 6-arm trial to examine the therapeutic effects of certain supplemental foods and as the outcome of his work, he was able to identify that sailors who ate oranges and lemons as part of their diet did not fall prey to scurvy. While it took some time to identify that these supportive features were due to the vitamin C in these citrus fruits [1], this first modern-day example of a clinical trial helped lay the foundation for how science and medicine has continued to evolve over the subsequent centuries. While this was considered a novel approach for assessing a potential treatment at the time, today it has been generally accepted that randomized clinical trials (RCTs) are the gold-standard for being able to assess a treatment for efficacy and safety.

There are of course many aspects to clinical trials, particularly statistical ones, and it is beyond the scope of this work to identify them all in great detail [2, 3], however, there are certain main points that are important to consider. Firstly, one of the main important features of RCTs is the randomization component, which allows for the random assignment of patients to different treatment groups. The goal of this is to help greatly reduce, and if possible eliminate, bias that can be introduced to the trial and affect the outcome of the results. For instance, an investigator may, either consciously or unconsciously, favor a certain type of patient into a treatment group or into a control group, which can negatively impact the viability and generalizability of a trial. For example, in the case where healthier patients are placed in the experimental treatment arm, it is possible that these patients have a more positive outcome at the end of the trial that is less likely to be due to the effectiveness of the treatment, but in part due to the overall healthier patients being enrolled in the treatment arms [4]. While it is true that randomization can help reduce the impact of confounding variables on a clinical trial, another potential way to help reduce this impact is through the use of stratified randomization, in which patients are stratified into treatment arms based on known factors that may influence the outcome response [5]. These stratification factors themselves are often identified in clinical literature through subsequent RCTs and studies in general that are able to elucidate which factors may be associated with certain clinical outcomes. For example, in the case of a diabetes trial, one may want to stratify patients into treatment arms based on baseline HbA1c, as it is known that this can have a high impact on treatment outcomes. This can be particularly useful in the case of smaller trials and may help with reducing sampling error [6]. Nevertheless, it is not always clear which variables may have an important impact on the outcomes of interest and thus it is here that randomization can help to limit imbalances across groups. It is for this reason the analyses presented in this work will focus primarily on data from RCTs.

However, even with randomization allocating the values of covariates theoretically to be equal amongst the arms of the trial, it is still possible that a differential heterogeneous subset of patients exists. Assessments of heterogeneous patient populations and outcomes allows researchers to examine if there are certain characteristics of a patient or group of patients that are important in differentiating overall outcomes. One common area this is seen in is within biomarkers, such as NT-proBNP, where the biomarker can be used as a surrogate for understanding cardiovascular disease improvement, progression, and severity [7]. Grouping patients by such biomarkers can in turn allow researchers and health care professionals with a better understanding of the likelihood of associated clinical outcomes for the respective subgroups [7].

Over the years, clinical trials have continued expanding upon the types of data they collect. Traditionally, clinical trials have focused their data collection to occur at individual time points throughout the trial at which the patient comes to the site. At these site visits, such data is collected as medical history data, laboratory data, blood tests and general wet biomarker data, clinical outcome data, and patient reported outcomes. Given the nature of some of these assessments, such as blood tests or scans, the patient needs to physically be at the site for the testing to occur. However, these visits are often few and far between throughout the course of the clinical trial and as a result the datapoints collected on each individual patient are often relatively sparse. This can be problematic for a few reasons depending on the type of data one is collecting. Variability often remains high due to the sparsity of the datapoints [8]. Particularly problematic are patient reported outcome data, in which one is often assessing subjective outcomes at a single point in time that can be influenced by a myriad of other factors in the patient's life.

Novel technologies are making headways into clinical trials whereby they provide complimentary information to data traditionally collected. In particular, wearables and biosensors allow for the passive and non-invasive collection of data with a higher resolution as compared to the more isolated data points currently collected and can also be used to collect data points over a long period of time. Furthermore, wearable and biosensor data can also provide us with a higher resolution of information about the changes developing within a patient over time. Moreover, wearables and biosensors are useful in the area of cardiovascular disease and heart failure as they can serve as proxies for measures of physical activity and quality of life. However, the most appropriate analytical approaches for examining these data are still under discussion.

1.2 Cardiovascular disease and heart failure

Cardiovascular disease and heart failure represent a public health crisis. For example, an estimated 5.7 million patients have been diagnosed with heart failure in the US with further increasing incidence and prevalence rates [9]. For patients diagnosed with heart failure, they experience frequent hospitalizations, a general decrease in quality of life, and increased risk for mortality [10]. It is important to note that heart failure carries a similar prognosis as many cancers with a 5-year survival rate of only 50 percent [11]. In addition to the burden placed upon the individual with a heart failure diagnosis, the treatment and care of individuals suffering from heart failure amounts to an enormous financial strain on healthcare systems. For example, Medicare spending in the United States allocates more funds annually to the treatment of heart failure than any other Medicare-covered condition [12]. Taken altogether, there is a clear need for further research and development efforts to focus in this area in an effort to help uncover treatments and therapeutics to ameliorate the effects of cardiovascular disease and heart failure on both the individual and society at large.

By expanding our therapeutic options for cardiovascular disease and heart failure, the scientific community can further reduce the impact on hospitals and the general health care system. Given the estimated incidence rate of 5.7 million heart failure patients in the US [9], earlier identification and intervention has the opportunity to improve healthcare outcomes and quality of life for patients. For example, given that actigraphy data can be collected with something like a cell phone that patients are likely already carrying around as part of their every day life, this data may be able to be used for classifying patients earlier before they begin to show clinical systems of heart failure. The median cost for heart failure hospitalization in the United States is \$ 13,418 and with close to 6 million patients needing hospitalizations [13], just the cost of hospitalizations can balloon to over \$ 80 billion dollars. Even in the case that 5% of patients are able to be identified prior to requiring a heart failure hospitalization, it could yield a potential savings of \$ 4 billion dollars.

While there are currently treatments available for heart failure [14], there is still room for additional novel therapeutics, in particular in the area of heart failure with preserved ejection fraction that has proven a difficult area to succeed in from a research perspective [15]. With the advent of new technologies entering the space, the use of wearables and biosensors offer a new modality from which to understand the disease development, patient journey, and impairments to quality of life for patients [16]. In this area of active research, biosensors and wearables are used to supplement data that is traditionally captured as part of a clinical trial. This work has applications to many therapeutic areas and indications such as neurology in the study of patients with Parkinson's disease or in the case of continuous examination of blood glucose levels in patients with diabetes [17]. Overall, these devices provide the utility of informing clinical practice by identifying and categorizing patients who represent with patterns of similarities or differences to one another. For this specific work, the focus will be primarily on the heart failure and cardiovascular space where there is a clear unmet medical need for improved therapeutics and a better understanding of the limitations. However, the methodologies and analytical approaches therein may apply to other indications as well.

Patients with heart failure have self-reported that some of the most important areas in which they are interested in improving include aspects related to quality-of-life improvements such as decreases in physical limitation [18]. This focus on patient reported outcomes has led many researchers in the area and development programs to examine subjective and objective measures of physical activity, such as the 6-minute walking test (6MWT) and the Kansas City Cardiomyopathy Questionnaire (KCCQ) [15]. In the case of the 6MWT, the patient attends a visit and walks as much as he/she can during a 6-minute time period and this distance is then recorded in meters, providing an objective measure of physical capability. However, there still is a lack of understanding of what level of activity patients feel comfortable engaging in when they are outside of the clinic. To assess the subjective aspects of cardiomyopathy, the KCCQ questionnaire is a 23-item instrument that is self-administered and quantifies such symptomology as physical functioning, social functioning, and general aspects of quality of life [19].

In order to administer these tests and capture both the objective and subjective aspects of cardiovascular functioning, the patient must attend an in person visit at the clinic. This has been the traditional method for how physical limitation has been assessed over the years. While a patient's physical capability as measured in a clinic is a useful measure to better understand a patient's capabilities, it provides one perspective of the patient's functioning. For example, the 6MWT is performed in an artificial capacity, within an artificial and controlled environment, that may not mimic that of the patient's true physical functioning characteristics. Conversely, with the data from wearables and biosensors, the researcher is able to better understand the daily activity profile of a patient and to better elucidate the pattern therein when a patient is in his/her own personal environment. From this type of data, it may be possible to answer such questions as 1) Over the course of time, does the patient increase or decrease their daily physical activity levels? 2) Is he/she able to sustain the activity for longer periods of time? In some ways, one can relate the relationship between the 6MWT and the additional data from wearables and biosensors to that of randomized clinical trials and real-world evidence. Undoubtedly RCTs provide a unique and important perspective to elucidate the underlying treatment effects for a potential therapeutic. However, over the last few decades it has become clear that understanding the patient in the real-world also serves a purpose. In a basic example, it may be easier to get a patient to follow a 5-times a day dosing regimen and assess the impact on disease progression as part of a clinical trial, particularly if as part of the participation in the clinical trial the patients receive some sort of notification system to ensure that they take the medicine. While allowing for a higher likelihood of adherence in the trial, examination of data in a real-world evidence scenario may provide a different picture that in fact patients may be having a difficult time adhering to these dosing schedules. Similarly, examining physical limitations as part of a clinical trial may prove different to an examination of physical activity data from wearables and biosensors in the comfort of a patient's own environment. In summary, it is important to not only capture what the patient is maximally capable of, which would be reflective in something like the 6MWT, but also the level of physical activity that a patient feels comfortable with engaging at home in the more real-world environment. This is particularly important given that patients may exhibit a differentiation in the level of physical activity they engage in during a clinic visit which reflects more of an artificial environment as compared to the true activity the patient may want to engage in while on their own. Taken together with the data collected in the clinic, the activity data collected by wearable devices can provide a more cohesive picture of the patient and his/her physical functioning limitations. Furthermore, in the context of patients in a clinical trial, the use of wearables and biosensors may allow for a better understanding of changes over time that may be due to the treatment effect of the new therapeutic. Recently the FDA has come out in support of the use of novel endpoints such as accelerometry in their recent heart failure guidance [20], suggesting that the field overall may be moving towards utilizing these devices not only in an exploratory manner to understand the patient journey and disease progression but also from a primary or secondary endpoint perspective.

One of the major questions still pending in the literature is what can be done with these wearable and biosensor device data and what questions can be answered with it? How can the data be utilized to improve our understanding of patients, their lives, and their journeys? The work here proposes some answers to these questions by examining the data utilizing clustering algorithms to group patients and assess any differences amongst patients with respect to physical activity and clinical outcomes. These assessments can be utilized to assess the relationship between physical activity and the clinical outcomes and also to detect if there are any signals of a heterogeneous response to clinical outcomes based on physical activity. This can also be helpful with respect to providing a more precision medicine-based approach where patients and their trajectories can be more accurately estimated as more is learned about the patients, their diseases, and the patient journey. Given the strong association between physical limitation and cardiovascular disease and more specifically heart failure, additional research is necessary to examine the relationship between improvement in physical capability and physical exercise and the associated improvement in clinical outcomes, including reduced hospitalizations and mortality, and a general improvement in understanding of the patients. The focus of this work will be on a further examination for some analytical options for the analysis of such data. To get a better understanding of what may be possible with these data, this work discusses potential approaches for classification and prediction of the physical activity data utilizing both unsupervised and supervised machine learning techniques.

Chapter 2

Datasets

2.1 Background on dataset

The data from the subsequent analyses is part of a recent multi-center, randomized, placebo-controlled, parallel group, double-blind dose-finding phase 2b heart failure trial with reduced ejection fraction trial sponsored by Bayer Pharmaceuticals. All clinical trial participants provided an informed consent for their voluntary enrollment in the trial. As part of the clinical trial, a wide variety of data from non-wearables and wearables, including but not limited to clinical biomarkers, medical history and demographics, and clinical outcomes data were collected. The patients enrolled included patients with a diagnosis of heart failure and a reduced ejection fraction, defined as a left ventricular ejection fraction $\leq$ 35 percent. Patients spent 20 weeks in the trial. All patients were 18 years of age or older at the time of enrollment in the trial.

2.2 Physical activity data from wearables and biosensors

Data related to activity intensity and activity duration were collected and derived from a chest-worn device that collected tri-axial accelerometry data (X/Y/Z axes) at 4-second intervals. Data were collected for the 7-days prior to randomization where each patient wore the device for the 7 consecutive days. These were considered the baseline values. Patients were selected out of the full data set who had data available for two wearable device variables: activity intensity (in milligravitational units (mg)) and activity duration (in seconds) [21]. In preparation for each wear periods, patients were given the device a day prior to their 7-day wear period and each patient had a subsequent scheduled visit at the clinic the day after the last day of the 7-day wear period for the device, which yielded a high level of adherence to the designated wear time. Data from each patient was assessed to ensure 7 full days of wear time, defined as wearing the device for at least 90 percent of the day for the full 7 days. For each patient, intra-day averages were calculated based on hourly aggregations for each variable and for each day such that each patient would end up with 14 total intra-day averages (7-days of intra-day averages for each of the two variables) which resulted in a cleaned data set representing approximately 85 percent of the total enrolled patients in the trial with respect to the baseline activity data. This data was used for the initial k-means clustering and the subsequent predictive models. Descriptive statistics were performed on the variables to assess for any potential issues resulting from outliers. Scatterplots of the individual variables were produced and leftward skewness was observed across the variables. Log transformations were applied to transform the data to allow for an approximately normal distribution. Standardization of the variables was performed such that the mean equaled 0 and the standard deviation was equal to 1.

2.3 Physical activity duration and intensity across the trial

A further subset of data was analyzed for patients who wore the device for 4 separate times throughout the trial (3 in addition to the earlier baseline assessment) for 7-days at each time for a total of 28 days. Patients were excluded from the analyses if they did not complete the trial (e.g. the patient dropped out of the trial or withdrew consent for further follow-up), and thus the patient did not have complete data for all 28 days. Complete data for a given patient required them to wear the device for at least 90 percent of the day for the full 28 days of wear time. For each patient, intra-day averages were calculated based on hourly aggregations for each variable such that a total of 56 intra-day averages were evaluated for each participant (28-days of intra-day averages computed for each of the two variables, yielding 56 averages per patient).

This dataset expands upon the previous dataset by containing a longitudinal assessment of patients and was only used in the consensus clustering portion of the analyses to assess the impact of examining the impact of a longitudinal assessment. Descriptive statistics were performed on the subsequent computed variables to assess for any issues such as skewness and outliers. Subsequent scatterplots were produced to examine the data and identified slight leftward skewness across the variables. To mitigate this in the subsequent analyses, a log transformation was applied to the data to transform it to an approximately normal distribution. The variables were then standardized with mean equal to 0 and standard deviation equal to 1. In preparation for the analysis, a PCA was performed and plotted to ensure that there were no obvious deviations from the assumptions, such as non-linear structures and/or cluster imbalance.

Descriptive statistics were performed on the variables to assess for any potential issues resulting from outliers. Scatterplots of the individual variables were produced and leftward skewness was observed across the variables. Log transformations were applied to transform the data to allow for an approximately normal distribution. As preparation for the clustering algorithm, standardization of the variables was performed such that the mean equaled 0 and the standard deviation was equal to 1.

2.4 Medical history and baseline characteristic variables

Certain continuous variables were grouped as categorical to obscure individual patient profiles and to limit any ability to identify individual patients. The following clinical variables were selected for inclusion into the models given their known impact in cardiovascular disease and heart failure. Patients were categorized into three groupings of age: <65 years old, 65-75 years old, and >75 years old. Age is often considered an important risk factor for cardiovascular disease given that with aging, the overall cardiac structure continues to deteriorate, which in turn leads to higher risk factors for developing heart failure [22]. Sex was included to provide additional demographic information regarding the patients and also given that there is data to suggest that rates of cardiovascular disease is increasing in women while rates remain somewhat the same in men [23]. Country and regional grouping is another variable that is included given that there are often differences in guideline directed medical therapy on how cardiovascular disease and heart failure is treated across various regions [24]. Given areas of enrollment for patients in the study, patients were grouped into three categories: Eastern Europe, North America, Western Europe and Israel.

N-terminal pro b-type natriuretic peptide (NT-proBNP) is a substance produced by the heart and has been established in the literature as a biomarker for cardiovascular disease, recent cardiovascular decompensation, and general cardiovascular outcomes [7]. Patients were grouped into two categories of NT-proBNP: less than or equal to the median NT-proBNP value of the study at baseline and greater than the median NT-proBNP value of the study. At enrollment in the trial, patients were classified based on the New York Heart Association (NYHA) Functional Classification [25], which classifies heart failure patients into four categories based on how limited they are while engaging in physical activity. Almost the entirety of patients in the trial presented fall into class II – IV. Class II patients are categorized as having a slight limitation in physical activity, with ordinary activity resulting in shortness of breath and fatigue. Class III patients have higher levels of limitation engaging in physical activity, with low levels of activity resulting in severe shortness of breath and fatigue. Class IV patients are unable to engage in physical activity without a level of discomfort and exhibit heart failure symptoms even while at rest [25]. Several variables were included with reference to patient's previous medical histories and diagnoses prior to enrolling in the clinical trial given their strong association with clinical outcomes in cardiovascular disease and heart failure, including prior histories of: 1) heart failure hospitalization [26], diabetes [27], atrial fibrillation [28], and hypertension [29]. These variables were coded as binary with the patient either having a previous history or diagnosis of the respective variable or not.

Estimated Glomerular Filtration Rate (EGFR) is a biomarker that indicates how well an individual's kidneys are filtering out extra waste and water from the blood into the urine. Given the intricate relationship between kidney function and cardiovascular disease, many patients with cardiovascular disease present with comorbidities of low EGFR values, kidney damage, or general chronic kidney disease [30]. Patients were grouped into two categories of EGFR: less than or equal to an EGFR value of 60 ml/min/1.73 m2 and those with an EGFR value greater than 60 ml/min/1.73 m2. An EGFR value of 60 ml/min/1.73 m2 was selected as the cut off point given that previous literature indicated higher risk for heart failure hospitalizations (adjusted HR: 1.21; 95% CI: 1.00–1.47), cardiovascular death (adjusted HR, 1.53; 95% CI: 1.23–1.91), and all-cause death (adjusted HR, 1.47; 95% CI, 1.24–1.76) for patients with EGFR values of less than 60 ml/min/1.73 m2 when adjusting for potential confounders [31]. Lastly, body mass index (BMI) was included given that higher values of BMI (greater than or equal to 35) have been found to be associated with higher risks of death in patients with chronic heart failure [32]. Patients were grouped into two categories of BMI: less than or equal to 30 and greater than 30. The same categorizations of these variables are used across the analyses presented within this work.

2.5 Clinical outcomes

Three different clinical outcomes were assessed as part of the subsequent predictive modelling analyses, which included the following: 1) time to first heart failure hospitalization, 2) time to the composite endpoint of first event of heart failure hospitalization or urgent heart failure visit, and 3) time to the composite endpoint of first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death. These three clinical outcomes were selected for assessment given that they are commonly used as the outcome endpoints of clinical trials and present clinically relevant information to both the patients and health care providers [33].

For the clinical outcomes and predictive analysis sections, a total sample size of 347 patients were analyzed. 49 patients experienced the clinical event of time to first heart failure hospitalization (14.1%) and 298 (85.9%) did not. 55 patients experienced the clinical event of time to the composite endpoint of first event of heart failure hospitalization or urgent heart failure visit (15.9%) and 292 (84.1%) did not. Lastly, 58 (16.7%) patients experienced the clinical event of time to the composite endpoint of first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death and 289 (83.3%) patients who did not. Given that a substantial amount of patients experienced these clinical outcomes, it was important to understand what the predictor variables were impacting the outcomes and what role activity duration and activity intensity played in them. An improved understanding of the association between these predictor variables and the clinical outcomes can help improve many aspects of the clinical trial landscape, including ensuring that the appropriate patient population is enrolled in a trial and to improve the accurate grouping of patients into risk categories based on the likelihood of developing the clinical outcomes of interest.

Chapter 3

Methods: K-Means Clustering

3.1 Aggregation of patients based on physical activity duration and intensity through K-means clustering

The goal of the following work was to assess whether the baseline clustering of activity intensity and activity duration data from a wearable device can be used to distinguish clinically relevant and unique phenotypes of patients and provide an improved understanding of patients enrolled in the trial. More specifically, physical activity and activity duration at baseline was clustered and the subsequent clustered groupings of patients were examined with respect to their demographics and medical histories to assess whether unique clinical phenotypes were present amongst the heterogeneous patient population. In this analysis, the K-means clustering algorithm was utilized.

3.2 Methods

The R package ’NbClust’ was used and, more specifically, the function NbClust was employed to assess 30 indices for determining the appropriate number of clusters by varying the distance measures [34]. The overwhelming majority of indices identified 3 clusters as the most appropriate under the k-means clustering method and such the results below were examined with these 3 clusters in mind. K-means clustering was then performed on the standardized data set with 3 pre-specified clusters with the added argument of 25 initial configurations attempted to converge on the optimal configuration. The subsequent clusters were then merged with a data set containing the baseline characteristics of the patients to examine the baseline characteristics of each cluster. This allowed assessment of potential numerical differences with respect to clinical characteristics and phenotypes. As these analyses were not multiplicity controlled, no hypothesis testing was performed.

3.3 Results

The R function NbClust from R package ‘NbClust’ was utilized to perform a grid search with 30 indices to identify the appropriate number of clusters for the data. This allows for an objective assessment for the appropriate number of clusters as compared to a-priori proposing a specific value without adequately assessing various indices to limit potential bias in the selected value. [34] include a full list of the indices. The grid search performed examined the 30 indices starting from k = 0 up to k = 15 utilizing the Euclidean distance measure to compute the dissimilarity matrix. The final convergence of the grid search yielded k=3 as the optimal value of clusters, as shown in Figure 3.1:

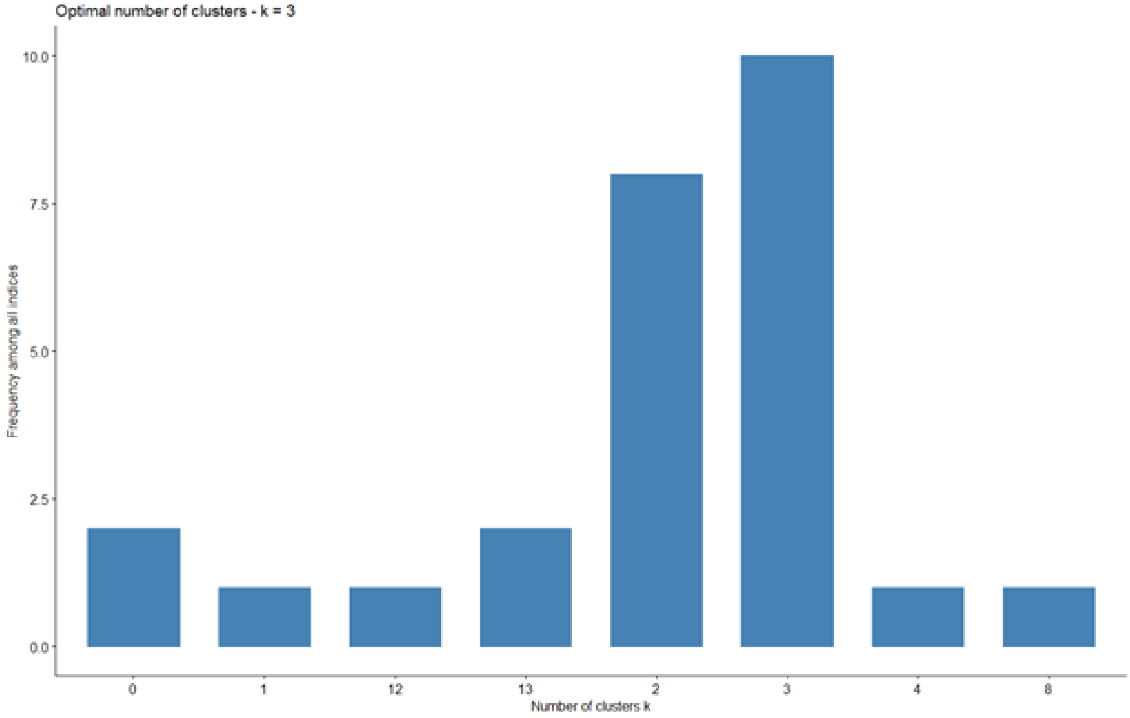


Figure 3.1: Plot visualizing optimal number of clusters based on the grid search optimization

With K identified as 3 based on the above grid search, the K-means clustering algorithm was run. A cluster plot was then created to assess the general separation and overlap amongst the clusters. Figure 3.2 indicates that there is a clear separation amongst the 3 clusters identified by the algorithm and that 84.1% of the cumulative variance is captured by the first two dimensions:

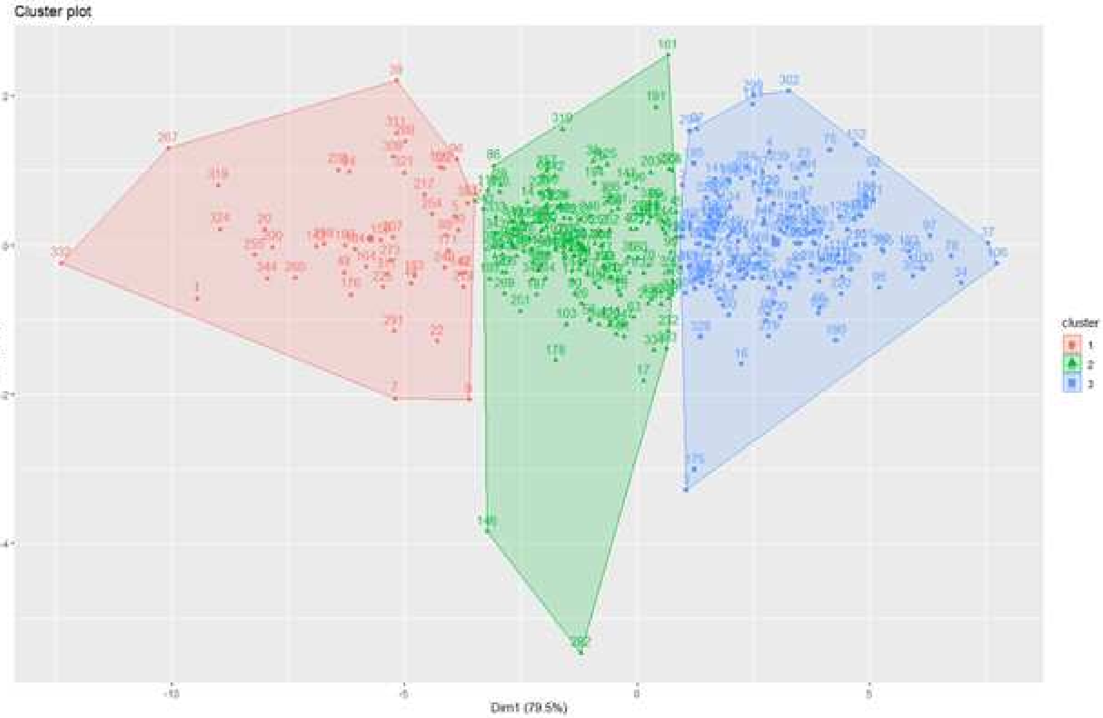
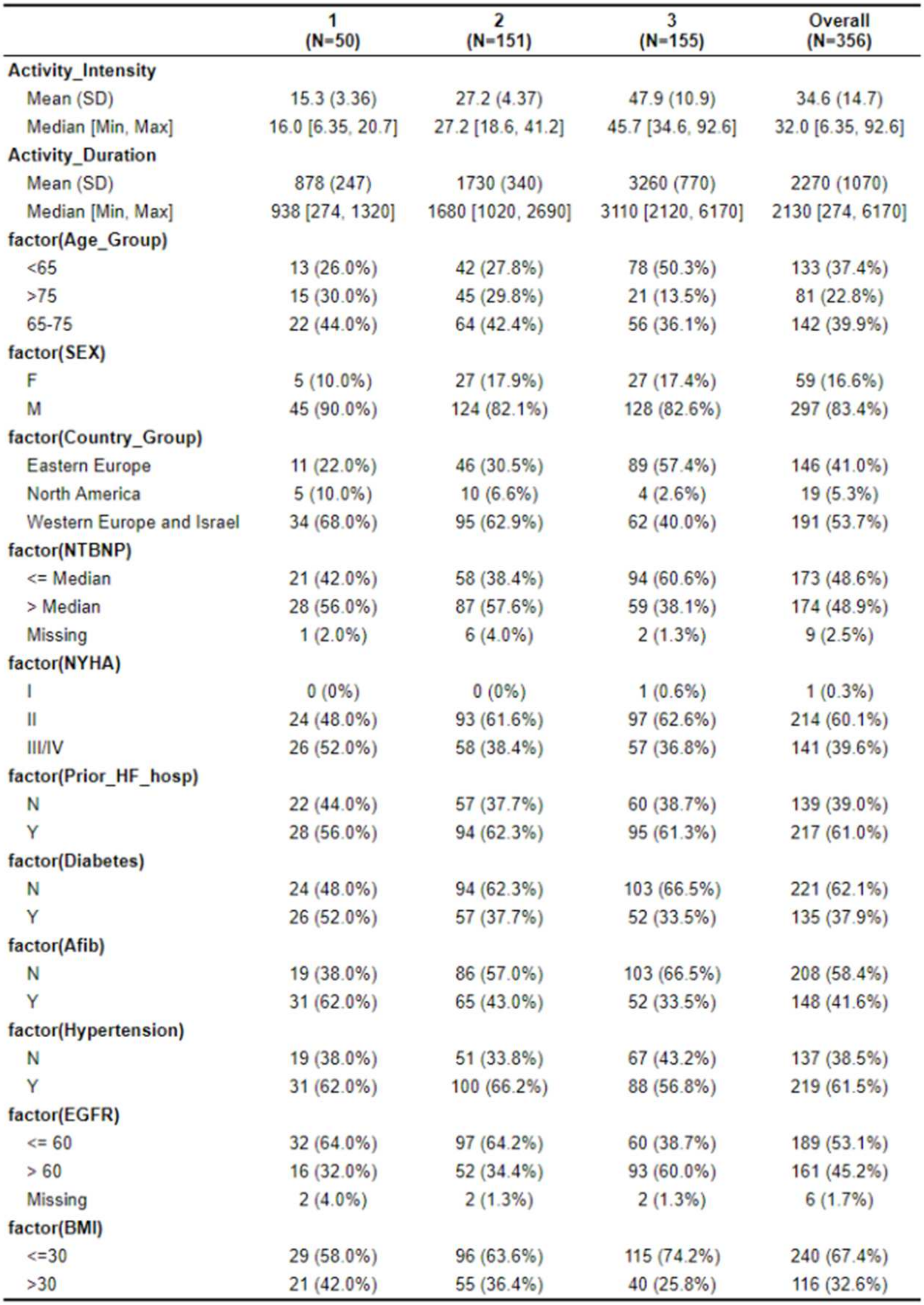


Figure 3.2: K-means patient cluster plot visualizing the distribution of patients into the cluster groupings

To better understand if patients could be differentiated based on these groupings, the three clusters were grouped together and further examined across baseline medical history and demographic variables, resulting in the results outlined in Table X below. Summary statistics were run on the original activity intensity and activity duration variables by cluster as well. Cluster 3 resulted in patients with a higher mean and median activity intensity and activity duration level, followed by cluster 2 and then cluster 1. Furthermore, cluster 3 had the youngest patients over, as compared to clusters 1 and 2 which were relatively similar in their age demographics. There were no large differences amongst gender between the groups, however, generalizability based on females may be difficult given the low proportion of female patients overall. Patients from cluster 3 were more likely to be from Eastern Europe, have a baseline NT-proBNP value less than or equal to the medium of the trial, have a higher baseline EGFR, and a lower BMI than patients from clusters 1 and 2.

Table 3.1: K-means clustering output with respect to baseline medical history and demographics



	1 (N=50)	2 (N=151)	3 (N=155)	Overall (N=356)
Activity_Intensity				
Mean (SD)	15.3 (3.36)	27.2 (4.37)	47.9 (10.9)	34.6 (14.7)
Median [Min, Max]	16.0 [6.35, 20.7]	27.2 [18.6, 41.2]	45.7 [34.6, 92.6]	32.0 [6.35, 92.6]
Activity_Duration				
Mean (SD)	878 (247)	1730 (340)	3260 (770)	2270 (1070)
Median [Min, Max]	938 [274, 1320]	1680 [1020, 2690]	3110 [2120, 6170]	2130 [274, 6170]
factor(Age_Group)				
<65	13 (26.0%)	42 (27.8%)	78 (50.3%)	133 (37.4%)
>75	15 (30.0%)	45 (29.8%)	21 (13.5%)	81 (22.8%)
65-75	22 (44.0%)	64 (42.4%)	56 (36.1%)	142 (39.9%)
factor(SEX)				
F	5 (10.0%)	27 (17.9%)	27 (17.4%)	59 (16.6%)
M	45 (90.0%)	124 (82.1%)	128 (82.6%)	297 (83.4%)
factor(Country_Group)				
Eastern Europe	11 (22.0%)	46 (30.5%)	89 (57.4%)	146 (41.0%)
North America	5 (10.0%)	10 (6.6%)	4 (2.6%)	19 (5.3%)
Western Europe and Israel	34 (68.0%)	95 (62.9%)	62 (40.0%)	191 (53.7%)
factor(NTBNP)				
<= Median	21 (42.0%)	58 (38.4%)	94 (60.6%)	173 (48.6%)
> Median	28 (56.0%)	87 (57.6%)	59 (38.1%)	174 (48.9%)
Missing	1 (2.0%)	6 (4.0%)	2 (1.3%)	9 (2.5%)
factor(NYHA)				
I	0 (0%)	0 (0%)	1 (0.6%)	1 (0.3%)
II	24 (48.0%)	93 (61.6%)	97 (62.6%)	214 (60.1%)
III/IV	26 (52.0%)	58 (38.4%)	57 (36.8%)	141 (39.6%)
factor(Prior_HF_hosp)				
N	22 (44.0%)	57 (37.7%)	60 (38.7%)	139 (39.0%)
Y	28 (56.0%)	94 (62.3%)	95 (61.3%)	217 (61.0%)
factor(Diabetes)				
N	24 (48.0%)	94 (62.3%)	103 (66.5%)	221 (62.1%)
Y	26 (52.0%)	57 (37.7%)	52 (33.5%)	135 (37.9%)
factor(Afib)				
N	19 (38.0%)	86 (57.0%)	103 (66.5%)	208 (58.4%)
Y	31 (62.0%)	65 (43.0%)	52 (33.5%)	148 (41.6%)
factor(Hypertension)				
N	19 (38.0%)	51 (33.8%)	67 (43.2%)	137 (38.5%)
Y	31 (62.0%)	100 (66.2%)	88 (56.8%)	219 (61.5%)
factor(EGFR)				
<= 60	32 (64.0%)	97 (64.2%)	60 (38.7%)	189 (53.1%)
> 60	16 (32.0%)	52 (34.4%)	93 (60.0%)	161 (45.2%)
Missing	2 (4.0%)	2 (1.3%)	2 (1.3%)	6 (1.7%)
factor(BMI)				
<=30	29 (58.0%)	96 (63.6%)	115 (74.2%)	240 (67.4%)
>30	21 (42.0%)	55 (36.4%)	40 (25.8%)	116 (32.6%)

3.4 Discussion and interpretation

Based on the results of the K-means clustering algorithm, there seem to be some unique clinical phenotypes present within the data. For example, cluster 3 is able to engage in the highest levels of physical activity intensity as well as for the longest duration, resulting in almost twice as long of a mean and median activity duration as compared to cluster 1 and almost four times the mean and median of activity duration of cluster 1. Furthermore, cluster 3 contains the youngest patients, with more than half of the cluster (50.3%) younger than the age of 65, compared to only 26.0% and 27.8% in the same age category in clusters 1 and 2, respectively. While there are no differences in gender across the clusters, there is a difference in country grouping with slightly more patients in cluster 3 enrolling from Eastern Europe as compared to only 22.0% and 30.5% from clusters 1 and 2, respectively.

As mentioned previously, the biomarker NT-proBNP can be used as a biomarker and proxy for cardiovascular disease and recent heart failure decompensation with higher values of NT-proBNP indicating a more recent decompensation as compared to lower values. Patients with higher values of NT-proBNP can be considered to be less stable than those with lower values of NT-proBNP and have a higher likelihood of worsening cardiovascular outcomes in the future [7]. With respect to NT-proBNP, 60.6% of patients from cluster 3 have a value less than or equal to the medium, as compared to only 42.0% and 38.4% in clusters 1 and 2, respectively. Patients from cluster 3 were also less likely to have a prior diagnosis of diabetes, a medical history of atrial fibrillation, or a medical history of hypertension as compared to clusters 1 and 2. EGFR is another variable for which the clusters differed. 60.0% of patients in cluster 3 had an EGFR value of greater than 60 ml/min/1.73 m2 as compared to only 32.0% and 34.4% in clusters 1 and 2. As previously discussed, EGFR values lower than 60 ml/min/1.73 m2 are associated with increased risks of heart failure hospitalization, all cause death, and cardiovascular death, and thus it is interesting to note that the cluster of patients who were able to engage in the highest level of activity intensity and the longest activity duration out of the groups of clusters also had the highest values of EGFR.

Lastly, cluster 3 contained 74.2% patients with a baseline BMI less than or equal to 30, compared to 58.0% and 63.6% in the same category in clusters 1 and 2, respectively.

Overall, clusters 1 and 2 may represent an older and potentially sicker heart failure patient population, with higher NT-proBNP values, lower EGFR, and higher rates of comorbidities. Furthermore, given cluster 1's lower baseline EGFR values and higher proportion of diabetes, there may be a higher prevalence of kidney disease involvement within this group. Cluster 3 may represent a younger and potentially separate heart failure etiology given the lower proportions of comorbidities and higher capability of engaging in longer and more strenuous bouts of activity duration and intensity. Interpreting the totality of the data, cluster 3 appears to present with the healthiest patients, followed by cluster 2 which acts as a sort of middle ground cluster that shows some further limitations based on activity intensity and activity duration profiles, however, not as highly limited as those patients belonging to cluster 1. Taken collectively, the results from these K-means clustering analyses indicate that clustering patients by activity intensity and activity duration data can be useful to assess baseline characteristics and differences amongst the clusters. This information can then be further used to generate hypotheses and for precision medicine purposes to identify if there are certain patients who may experience different clinical outcomes. Furthermore, the results suggest that while many important variables are considered as part of traditional clinical trials, the information gained from wearables and biosensors are complementary and have the potential to help the scientific community better understand the patient journey and clinical phenotypes of patients. Coupled with traditional data, these complementary approaches can be used to further refine research and development programs in their aims of identifying novel therapeutics and cures for today's most pervasive and pernicious diseases.

Chapter 4

Methods: Consensus Clustering

4.1 Aggregation of patients based on physical activity duration and intensity through Consensus clustering

One of the major limitations of the K-means clustering algorithm is the limitation of repeated runs of the algorithm yielding different results for patient clusters. Patient A may fall into cluster 1 during run 1 of the K-means algorithm, however, his/her membership may fall into cluster 2 during run 2, and the clusters themselves are likely to be produced from different convergences of the K-means algorithm. In an attempt to rectify this, a consensus clustering approach is utilized to address the following main concerns: 1) the resulting clusters are dependent on an arbitrary distance selection, variability of clustering results based on the initial clustering selections, the overall difficulty in validating and generalizing clustering results ([35, 36]. For consensus clustering, the R package "M3C" is utilized and, specifically, a partitioning around medoids (PAM) clustering algorithm with Euclidean distance and an entropy objective function is examined. A grid search is used to identify the appropriate value of K and then the clusters were examined with respect to their baseline characteristics, medical history, and activity duration and intensity. Unique clinical phenotypes were identified and the utility of these approaches to identify possible clusters of heterogeneous patient populations is discussed.

4.2 Methods

The R package 'M3C' from the Bioconductor suite within R was used to perform the Monte Carlo Reference-based Consensus Clustering algorithm [37]. This package utilizes a Monte Carlo simulation approach which keeps the original correlational structures of the input data, resulting in the creation of multivariate Gaussian references to aid in the consensus clustering and subsequently reduce intrinsic bias. More specifically, a partition around medoids (PAM) clustering algorithm with Euclidean distance and an entropy objective function was utilized, in which the consensus matrix elements are treated as probabilities. In this approach, the algorithm utilizes the information entropy and attempts to minimize it to find the corresponding most appropriate value of K. In such an approach, the K corresponding to the minimized information entropy is indicative of more stability as well as less uncertainty within the system throughout the iterative resampling of the consensus clustering algorithm. As an output to these analyses, PAC scores by cluster are initially examined, which identify the consensus matrix stability under each value of K. However, given the common score bias favoring lower values of K is present within the PAC scoring, the relative cluster stability index (RCSI) is examined with an attempt to maximize the function given that this eliminates the score bias [37]. This is then combined with a resulting p-value to select an appropriate value of K and assess whether the null hypothesis that K=1 should be rejected in favor of a value of K > 1 indicating that a true clustering exists within the data. Furthermore, the M3C package itself provides an output of the K value it selects as most appropriate.

4.3 Results

Several different outputs were assessed to identify the appropriate value of K. Figure 1 presents the RCSI graph indicating the appropriateness of selecting K=4 as the number of clusters given that this is the value of K for which the function is maximized and suggests the highest stability index at a value of K=4. In addition to the RCSI, the p-value assessing whether it is appropriate to reject the null hypothesis of K=1 is significant at the K=4 value. Thus, we reject the null hypothesis that K=1 in favor of the K=4 value (Monte Carlo simulation p-value = 0.038 and normalized p-value = 0.00004). Figure 4.1 below shows the RCSI across K clusters plot identifying K=4 as the appropriate value of K for which RCSI is maximized.

Figure 4.2 below examines the information entropy graph indicating the appropriateness of selecting K=4 as the number of clusters given that this is the value of K for which the function is minimized and thus indicative of more stability throughout the iterative sampling across the consensus clustering algorithm at K=4.

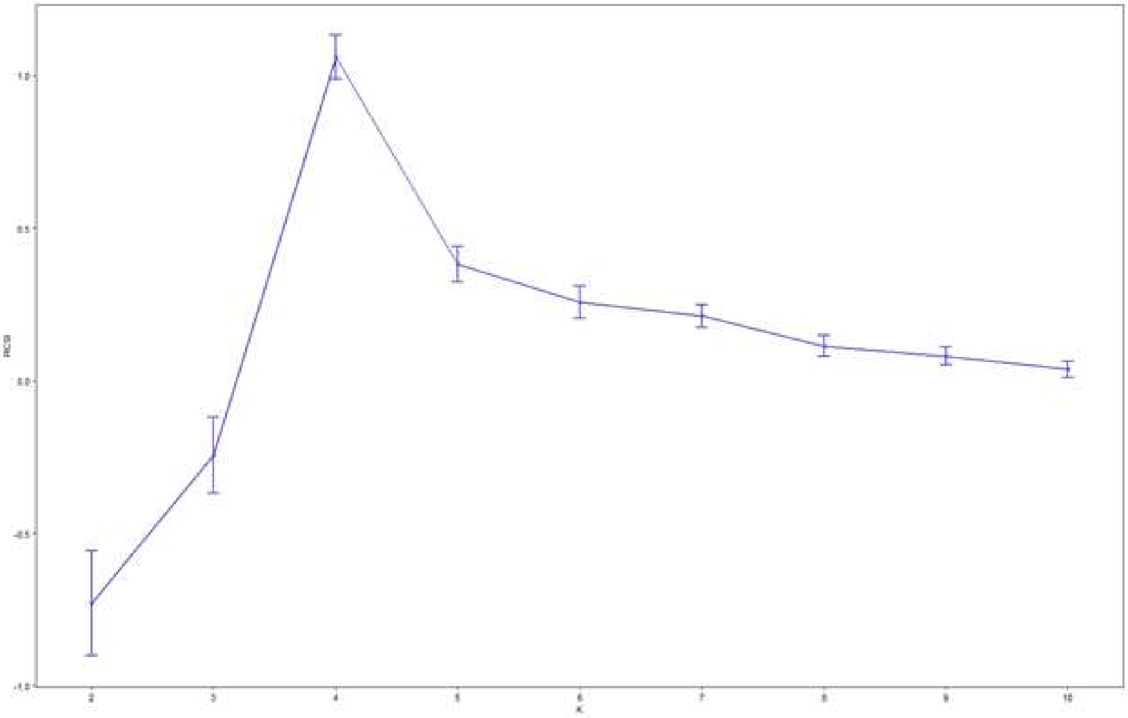


Figure 4.1: RCSI plot visualizing the RCSI across K clusters

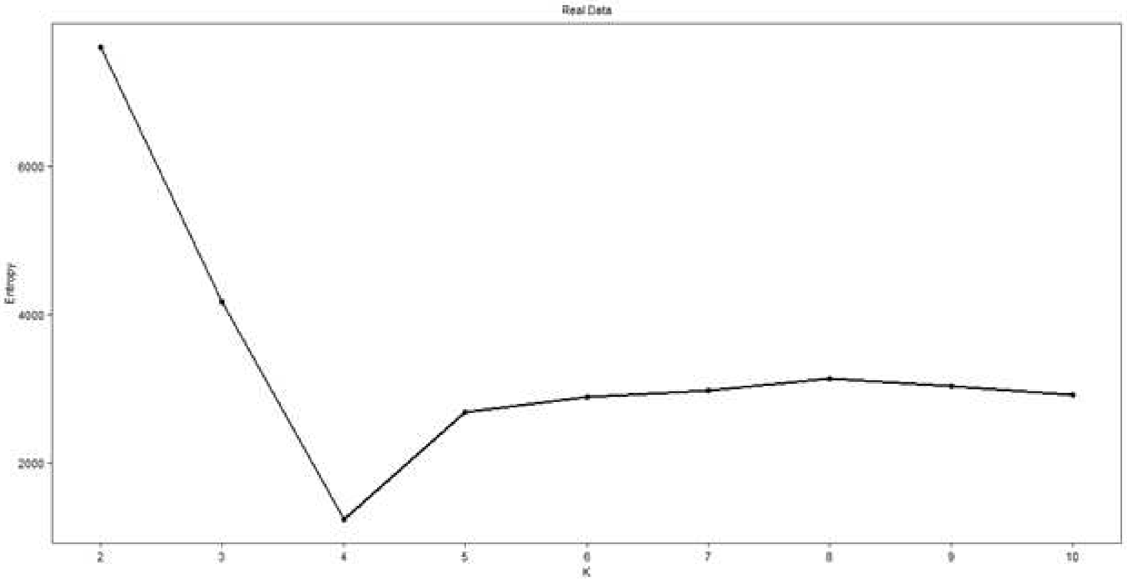
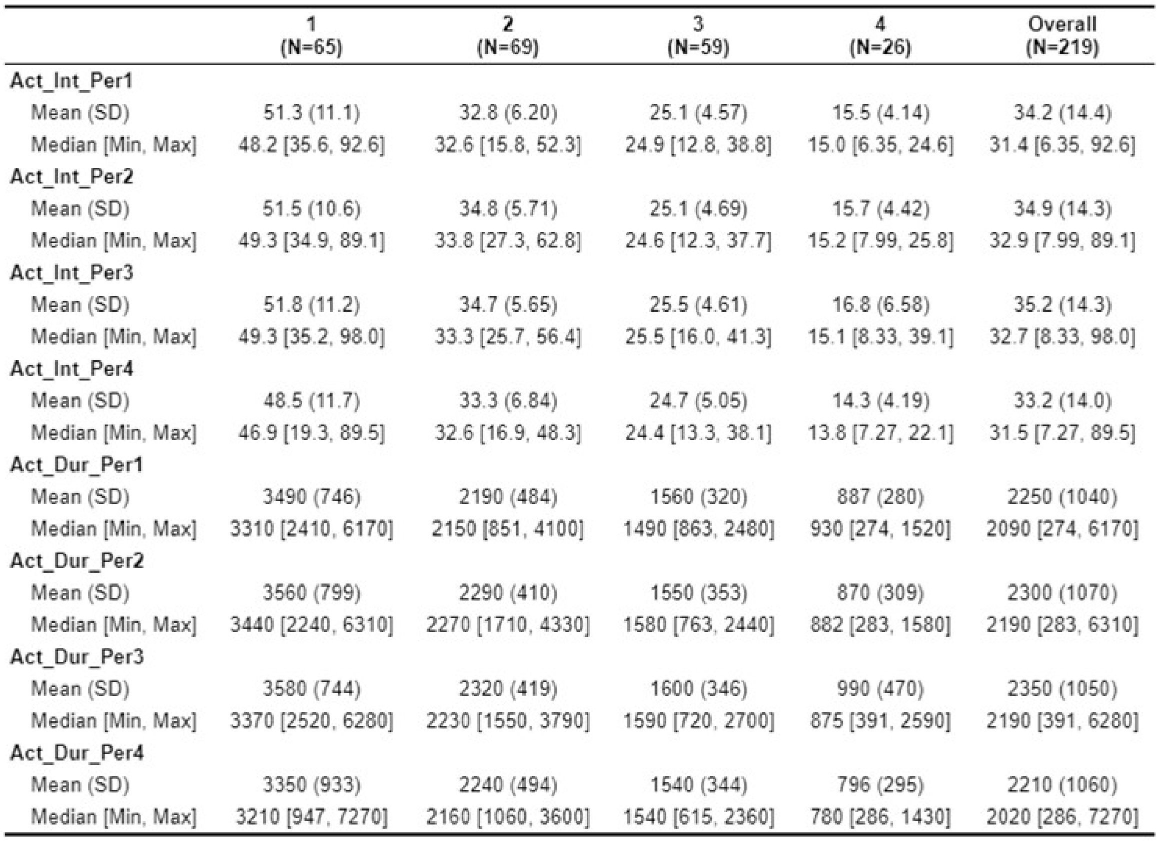


Figure 4.2: Entropy plot visualizing the information entropy graph

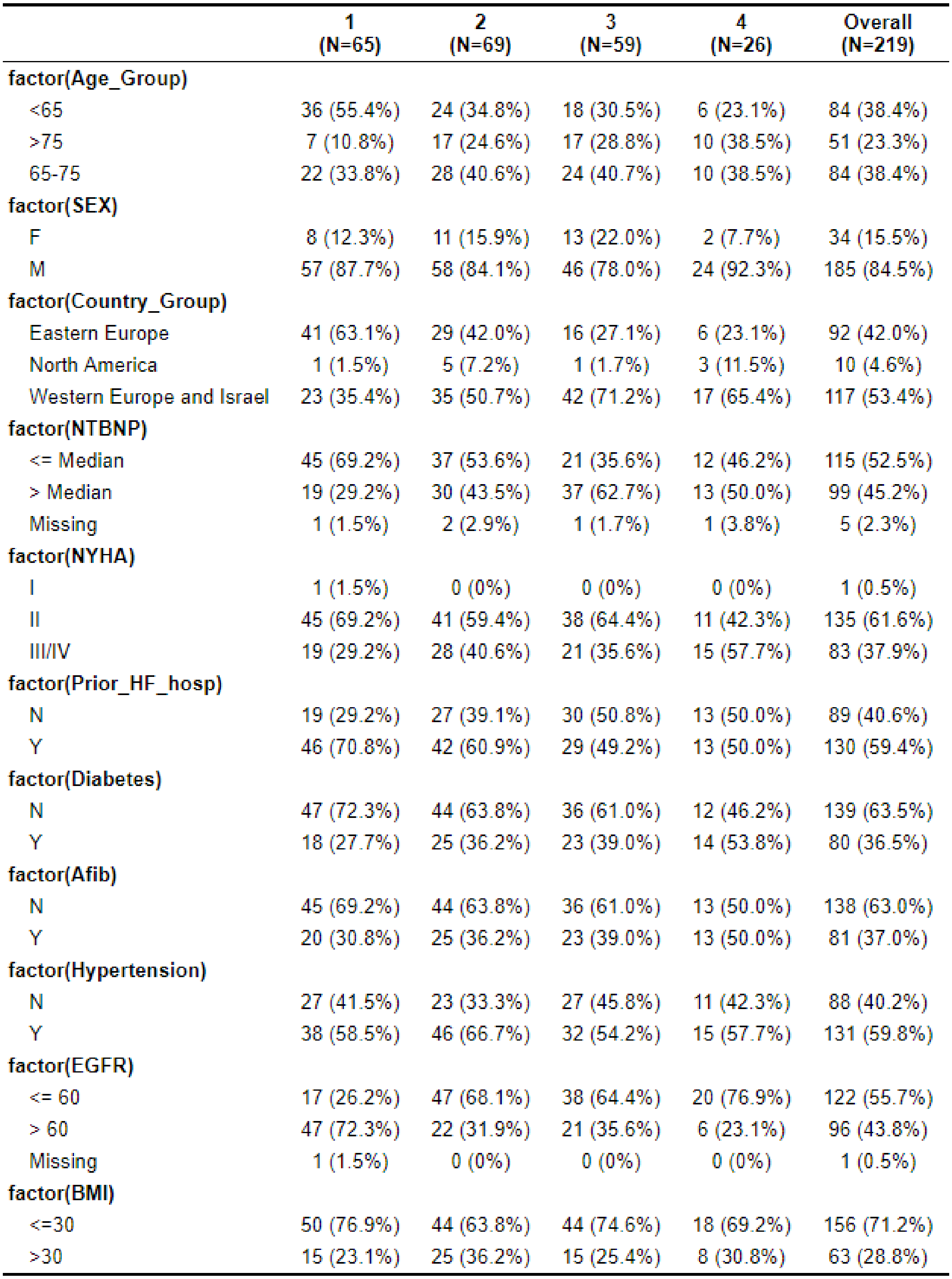
Furthermore, the M3C package itself provides an assessment of which value of K is the most appropriate given the output of several assessments including those discussed above. In the analyses examined here, the M3C algorithm selected K=4 as the appropriate K value as well. Thus, the overall criteria examination identified 4 clusters as the most appropriate for the data under this approach. Data from these subsequent clusters was then merged with baseline medical history and demographic data to examine characteristics of each cluster as well as differences in activity intensity and duration (Tables Table 4.1 and Table 4.2 below). Given that these analyses were not multiplicity controlled, no subsequent hypothesis testing was performed and the subsequent examinations focus on identifying numerical differences and the potential elucidation of clinical characteristics suggesting unique clinical phenotypes within the data.

Table 4.1: Descriptive statistics of activity intensity and duration across clusters identified by consensus clustering



	1 (N=65)	2 (N=69)	3 (N=59)	4 (N=26)	Overall (N=219)
Act_Int_Per1					
Mean (SD)	51.3 (11.1)	32.8 (6.20)	25.1 (4.57)	15.5 (4.14)	34.2 (14.4)
Median [Min, Max]	48.2 [35.6, 92.6]	32.6 [15.8, 52.3]	24.9 [12.8, 38.8]	15.0 [6.35, 24.6]	31.4 [6.35, 92.6]
Act_Int_Per2					
Mean (SD)	51.5 (10.6)	34.8 (5.71)	25.1 (4.69)	15.7 (4.42)	34.9 (14.3)
Median [Min, Max]	49.3 [34.9, 89.1]	33.8 [27.3, 62.8]	24.6 [12.3, 37.7]	15.2 [7.99, 25.8]	32.9 [7.99, 89.1]
Act_Int_Per3					
Mean (SD)	51.8 (11.2)	34.7 (5.65)	25.5 (4.61)	16.8 (6.58)	35.2 (14.3)
Median [Min, Max]	49.3 [35.2, 98.0]	33.3 [25.7, 56.4]	25.5 [16.0, 41.3]	15.1 [8.33, 39.1]	32.7 [8.33, 98.0]
Act_Int_Per4					
Mean (SD)	48.5 (11.7)	33.3 (6.84)	24.7 (5.05)	14.3 (4.19)	33.2 (14.0)
Median [Min, Max]	46.9 [19.3, 89.5]	32.6 [16.9, 48.3]	24.4 [13.3, 38.1]	13.8 [7.27, 22.1]	31.5 [7.27, 89.5]
Act_Dur_Per1					
Mean (SD)	3490 (746)	2190 (484)	1560 (320)	887 (280)	2250 (1040)
Median [Min, Max]	3310 [2410, 6170]	2150 [851, 4100]	1490 [863, 2480]	930 [274, 1520]	2090 [274, 6170]
Act_Dur_Per2					
Mean (SD)	3560 (799)	2290 (410)	1550 (353)	870 (309)	2300 (1070)
Median [Min, Max]	3440 [2240, 6310]	2270 [1710, 4330]	1580 [763, 2440]	882 [283, 1580]	2190 [283, 6310]
Act_Dur_Per3					
Mean (SD)	3580 (744)	2320 (419)	1600 (346)	990 (470)	2350 (1050)
Median [Min, Max]	3370 [2520, 6280]	2230 [1550, 3790]	1590 [720, 2700]	875 [391, 2590]	2190 [391, 6280]
Act_Dur_Per4					
Mean (SD)	3350 (933)	2240 (494)	1540 (344)	796 (295)	2210 (1060)
Median [Min, Max]	3210 [947, 7270]	2160 [1060, 3600]	1540 [615, 2360]	780 [286, 1430]	2020 [286, 7270]

Table 4.2: Baseline and demographic data across clusters identified by consensus clustering



	1 (N=65)	2 (N=69)	3 (N=59)	4 (N=26)	Overall (N=219)
factor(Age_Group)					
<65	36 (55.4%)	24 (34.8%)	18 (30.5%)	6 (23.1%)	84 (38.4%)
>75	7 (10.8%)	17 (24.6%)	17 (28.8%)	10 (38.5%)	51 (23.3%)
65-75	22 (33.8%)	28 (40.6%)	24 (40.7%)	10 (38.5%)	84 (38.4%)
factor(SEX)					
F	8 (12.3%)	11 (15.9%)	13 (22.0%)	2 (7.7%)	34 (15.5%)
M	57 (87.7%)	58 (84.1%)	46 (78.0%)	24 (92.3%)	185 (84.5%)
factor(Country_Group)					
Eastern Europe	41 (63.1%)	29 (42.0%)	16 (27.1%)	6 (23.1%)	92 (42.0%)
North America	1 (1.5%)	5 (7.2%)	1 (1.7%)	3 (11.5%)	10 (4.6%)
Western Europe and Israel	23 (35.4%)	35 (50.7%)	42 (71.2%)	17 (65.4%)	117 (53.4%)
factor(NTBNP)					
<= Median	45 (69.2%)	37 (53.6%)	21 (35.6%)	12 (46.2%)	115 (52.5%)
> Median	19 (29.2%)	30 (43.5%)	37 (62.7%)	13 (50.0%)	99 (45.2%)
Missing	1 (1.5%)	2 (2.9%)	1 (1.7%)	1 (3.8%)	5 (2.3%)
factor(NYHA)					
I	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)	1 (0.5%)
II	45 (69.2%)	41 (59.4%)	38 (64.4%)	11 (42.3%)	135 (61.6%)
III/IV	19 (29.2%)	28 (40.6%)	21 (35.6%)	15 (57.7%)	83 (37.9%)
factor(Prior_HF_hosp)					
N	19 (29.2%)	27 (39.1%)	30 (50.8%)	13 (50.0%)	89 (40.6%)
Y	46 (70.8%)	42 (60.9%)	29 (49.2%)	13 (50.0%)	130 (59.4%)
factor(Diabetes)					
N	47 (72.3%)	44 (63.8%)	36 (61.0%)	12 (46.2%)	139 (63.5%)
Y	18 (27.7%)	25 (36.2%)	23 (39.0%)	14 (53.8%)	80 (36.5%)
factor(Afib)					
N	45 (69.2%)	44 (63.8%)	36 (61.0%)	13 (50.0%)	138 (63.0%)
Y	20 (30.8%)	25 (36.2%)	23 (39.0%)	13 (50.0%)	81 (37.0%)
factor(Hypertension)					
N	27 (41.5%)	23 (33.3%)	27 (45.8%)	11 (42.3%)	88 (40.2%)
Y	38 (58.5%)	46 (66.7%)	32 (54.2%)	15 (57.7%)	131 (59.8%)
factor(EGFR)					
<= 60	17 (26.2%)	47 (68.1%)	38 (64.4%)	20 (76.9%)	122 (55.7%)
> 60	47 (72.3%)	22 (31.9%)	21 (35.6%)	6 (23.1%)	96 (43.8%)
Missing	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)	1 (0.5%)
factor(BMI)					
<=30	50 (76.9%)	44 (63.8%)	44 (74.6%)	18 (69.2%)	156 (71.2%)
>30	15 (23.1%)	25 (36.2%)	15 (25.4%)	8 (30.8%)	63 (28.8%)

Patients were assigned to the 4 identified clusters and then examined based on baseline characteristics and demographics. Furthermore, summary statistics were computed on the original activity duration and intensity variables for each wear period and then described by cluster. Numerical differences were identified. Across all periods of wear time, cluster 1 had numerically higher values of activity intensity and duration. This was followed by cluster 2 which often yielded activity intensity and duration values similar to that of the average across all clusters. Cluster 3 performed slightly worse than the overall average, followed by cluster 4 which performed numerically much worse than the other clusters. For example, cluster 1 had a 3.3-fold increase in mean activity intensity at wear period 1 and a 3.4-fold increase in mean activity intensity at wear period 4 compared to cluster 4. These results argue that the consensus clustering algorithm corrected clustered patients into 4 clusters that differed based on their activity duration and intensity profiles. These numerical differences were also evident when comparing clinical features (see Table 4.2). Cluster 1 presented with younger patients who were predominantly from Eastern Europe, a higher proportion of patients with baseline NT-proBNP values less than or equal to the median, a higher proportion of patients in the less severe NYHA classes (class II), a higher proportion of patients with previous heart failure hospitalization, a lower proportion of patients with diabetes and atrial fibrillation, a higher proportion of patients with larger baseline EGFR values, and a larger proportion of patients with lower BMI values as compared to the more severe cluster 4. Overall, the clinical features for patients in cluster 1 suggest that they are generally younger and healthier, however, it is interesting to note the elevated proportion of previous heart failure hospitalizations as compared to other clusters, perhaps indicating that they had begun some sort of stabilization therapy for their heart failure and thus presented as less severe and with more physical activity capacity than the patients from other clusters, and in particular cluster 4. Furthermore, while cluster 4 seemingly represented the oldest and least healthy patients, the results suggest that these patients comprising cluster 4 may have more renal impairment along with elevated rates of diabetes, hypertension, and atrial fibrillation. In general, patients from clusters 2 and 3 seem to represent patients comprising of moderate severity, having numerically higher proportions of diabetes, atrial fibrillation, and hypertension as compared to patients in cluster 1 but not as high as those in cluster 4.

4.4 Discussion: Limitations of clustering approach

There are some limitations to these approaches. As previously stated, statistical hypothesis testing was not performed given that these comparisons are not multiplicity adjusted and one would not have the appropriate power to detect differences amongst the groups. As these comparisons were not pre-specified and are post-hoc in nature, it is possible that they are detecting sample fluctuations and noise instead of actual clinical phenotypes within the patients. Furthermore, small numbers of patients are represented in cluster 4, and thus one may need to be careful with any general interpretations from these findings. All analyses here must be taken into context with biological plausibility and further assessment to be generalized to the larger population. Nevertheless, it is interesting to note that while these assessments were all conducted post-hoc, the clustering algorithms were only applied to the wearable device activity intensity and duration data, and yet yielded clusters that were clinically distinct from one another when examining baseline characteristics. Furthermore, these relationships within clusters suggested biological plausibility for the interpretation that patients who reached the highest levels of activity intensity and activity duration were the patients that presented with the least severe clinical features. Similar analyses can be used in the future to potentially identify differences in clinical phenotypes based on wearable device data and to help in the development of precision medicine therapeutics that are more tailored to clusters of clinical phenotypes that may be present within a sample of patients in an indication.

Chapter 5

Methods: Cox Proportional Hazards Model

5.1 Introduction to predictive models of clinical outcomes

As earlier outlined, the second step of these analyses is to examine the predictive relationship between the variables collected and those of clinical outcomes. The earlier presented clustering algorithms fall into the category of unsupervised learning wherein the algorithm does not train itself based on labeled output data. Instead, it functions by aggregating datapoints based on their relationship to one another from the unlabeled data [38]. The following predictive models examine the relationship between the variables and their impact on several clinical outcomes of interest. In particular, there are three sets of clinical outcomes that are assessed: 1) time to first heart failure hospitalization, 2) time to first heart failure hospitalization or urgent heart failure visit, 3) and for the composite of time to first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death.

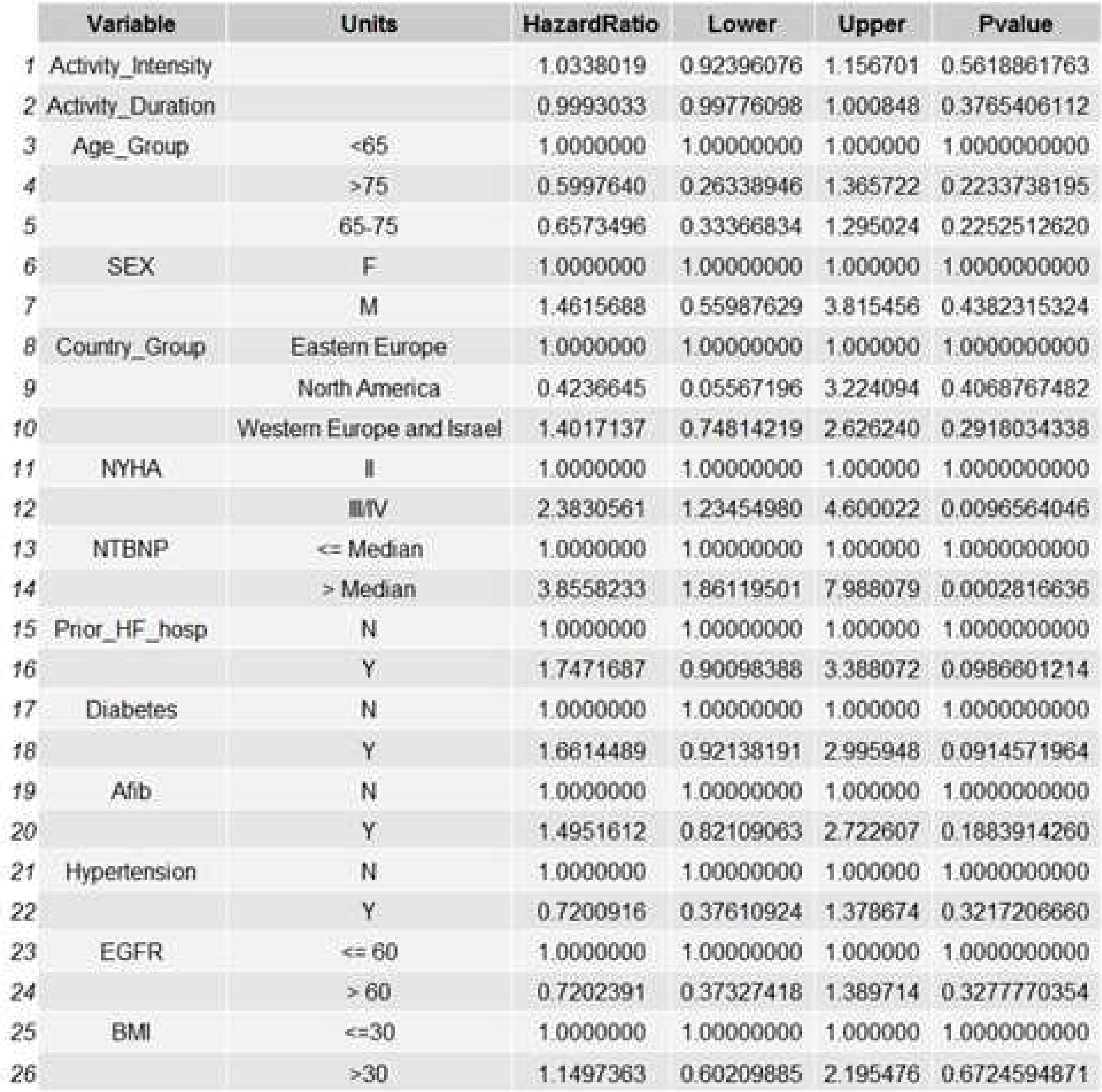
5.2 Cox Proportional Hazards Model and Method

In the first set of analyses, in order to examine the relationship between the physical activity variables and other data collected to the clinical outcomes, a Cox proportional hazards (PH) model was used [39]. The Cox PH model functions essentially as a regression model with the benefit that it can take time-to-event information, such as survival times of patients with respect to a clinical outcome, and assess the association between these outcomes and predictor variables. Bradburn and colleagues provide more theoretical information regarding the Cox PH [40]. In addition, a backwards selection approach to the Cox PH model is also fitted to the data. Utilizing the fastbw function from the R package rms, a fast backward elimination of the predictor variables is performed using the method from Lawless and Singhal [41]. Wald statistics and subsequent p-values are computed. The final output contains the predictive factors for the reduced model that are deemed to have a significant association with the clinical outcomes [42].

5.3 Results for time to first heart failure hospitalization based on Cox PH model

A Cox PH model was fitted to the data with the variables activity intensity, activity duration, sex, country grouping, NYHA class at baseline, NT-proBNP at baseline, prior heart failure hospitalization, diabetes, atrial fibrillation, hypertension, EGFR, and BMI as independent variables and the time-to-event clinical outcome of heart failure hospitalization as the dependent variable which yielded the following results:

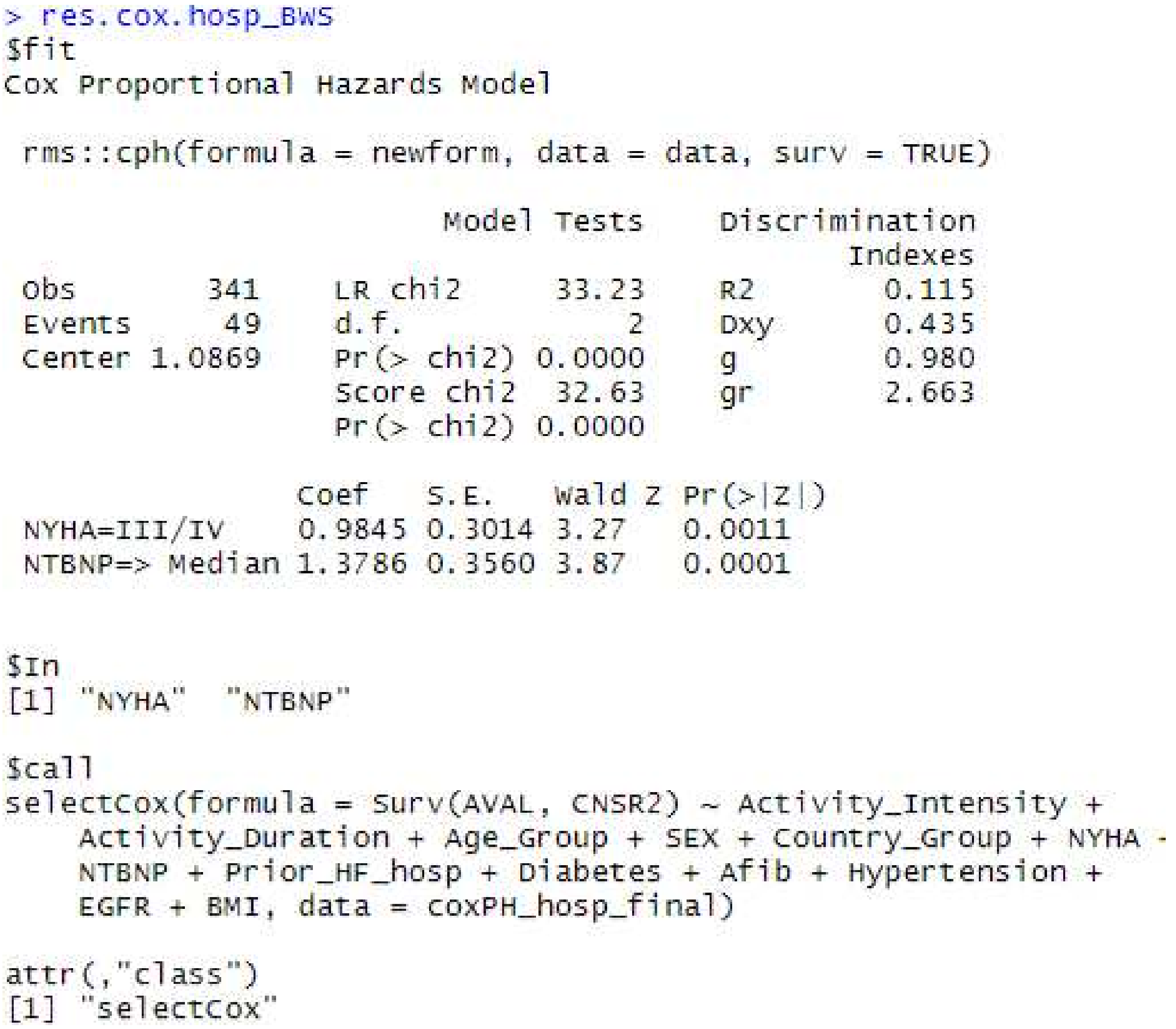
Table 5.1: Cox Proportional Hazards model output for time to first heart failure hospitalization



	Variable	Units	HazardRatio	Lower	Upper	Pvalue
1	Activity_Intensity		1.0338019	0.92396076	1.156701	0.5618861763
2	Activity_Duration		0.9993033	0.99776098	1.000848	0.3765406112
3	Age_Group	<65	1.0000000	1.00000000	1.000000	1.0000000000
4		>75	0.5997640	0.26338946	1.365722	0.2233738195
5		65-75	0.6573496	0.33366834	1.295024	0.2252512620
6	SEX	F	1.0000000	1.00000000	1.000000	1.0000000000
7		M	1.4615688	0.55987629	3.815456	0.4382315324
8	Country_Group	Eastern Europe	1.0000000	1.00000000	1.000000	1.0000000000
9		North America	0.4236645	0.05567196	3.224094	0.4068767482
10		Western Europe and Israel	1.4017137	0.74814219	2.626240	0.2918034338
11	NYHA	II	1.0000000	1.00000000	1.000000	1.0000000000
12		III/IV	2.3830561	1.23454980	4.600022	0.0096564046
13	NTBNP	<= Median	1.0000000	1.00000000	1.000000	1.0000000000
14		> Median	3.8558233	1.86119501	7.988079	0.0002816636
15	Prior_HF_hosp	N	1.0000000	1.00000000	1.000000	1.0000000000
16		Y	1.7471687	0.90098388	3.388072	0.0986601214
17	Diabetes	N	1.0000000	1.00000000	1.000000	1.0000000000
18		Y	1.6614489	0.92138191	2.995948	0.0914571964
19	Afib	N	1.0000000	1.00000000	1.000000	1.0000000000
20		Y	1.4951612	0.82109063	2.722607	0.1883914260
21	Hypertension	N	1.0000000	1.00000000	1.000000	1.0000000000
22		Y	0.7200916	0.37610924	1.378674	0.3217206660
23	EGFR	<= 60	1.0000000	1.00000000	1.000000	1.0000000000
24		> 60	0.7202391	0.37327418	1.389714	0.3277770354
25	BMI	<=30	1.0000000	1.00000000	1.000000	1.0000000000
26		>30	1.1497363	0.60209885	2.195476	0.6724594871

The above output yielded NYHA class and NTproBNP as the most important variables with respect to the clinical outcome of time to first heart failure hospitalization, as indicated by both variables having a p-value ≤ 0.05. The variable diabetes is trending towards significance, with a p-value ≤ 0.10. A backwards selection approach to the Cox PH model was also fitted to the data yielding the following results:

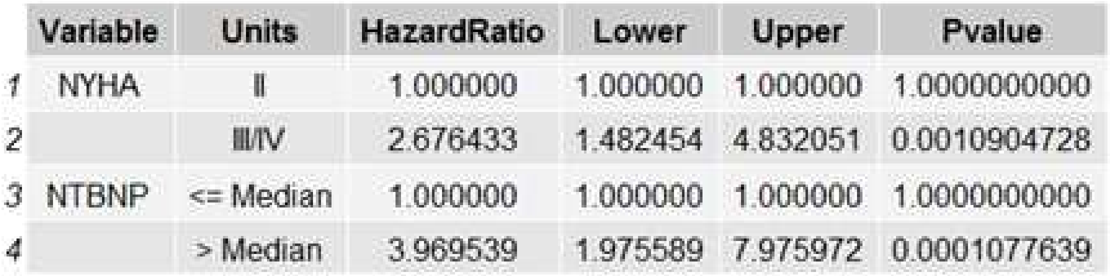
Table 5.2: Cox Proportional Hazards model with backwards selection output for time to first heart failure hospitalization



```
> res.cox.hosp_BWS
$fit
Cox Proportional Hazards Model

 rms::cph(formula = newform, data = data, surv = TRUE)

                        Model Tests    Discrimination
                                          Indexes
 Obs        341    LR chi2     33.23    R2       0.115
 Events      49    d.f.            2    Dxy      0.435
 Center 1.0869     Pr(> chi2) 0.0000    g        0.980
                   Score chi2  32.63    gr       2.663
                   Pr(> chi2) 0.0000

                Coef   S.E.   Wald Z Pr(>|Z|)
 NYHA=III/IV    0.9845 0.3014 3.27   0.0011
 NTBNP=> Median 1.3786 0.3560 3.87   0.0001


$In
[1] "NYHA"  "NTBNP"

$call
selectCox(formula = Surv(AVAL, CNSR2) ~ Activity_Intensity +
    Activity_Duration + Age_Group + SEX + Country_Group + NYHA -
    NTBNP + Prior_HF_hosp + Diabetes + Afib + Hypertension +
    EGFR + BMI, data = coxPH_hosp_final)

attr(,"class")
[1] "selectCox"
```

The backward selection approach further confirmed NYHA and NTproBNP as the only significant variables with respect to the outcome of time to first heart failure hospitalization. The Cox PH model was rerun to just include the variables NYHA and NTproBNP to assess the model features as shown in the following table:

Table 5.3: Cox Proportional Hazards model final reduced output for time to first heart failure hospitalization



	Variable	Units	HazardRatio	Lower	Upper	Pvalue
1	NYHA	II	1.000000	1.000000	1.000000	1.0000000000
2		III/IV	2.676433	1.482454	4.832051	0.0010904728
3	NTBNP	<= Median	1.000000	1.000000	1.000000	1.0000000000
4		> Median	3.969539	1.975589	7.975972	0.0001077639

Both variables remained significant in the final reduced model suggesting these are the appropriate predictors to include. To further assess the appropriateness of the Cox PH model with this data, an assessment of the underlying assumptions of the Cox PH model were performed by examining the Schoenfeld residuals as shown in the figure below [43]:

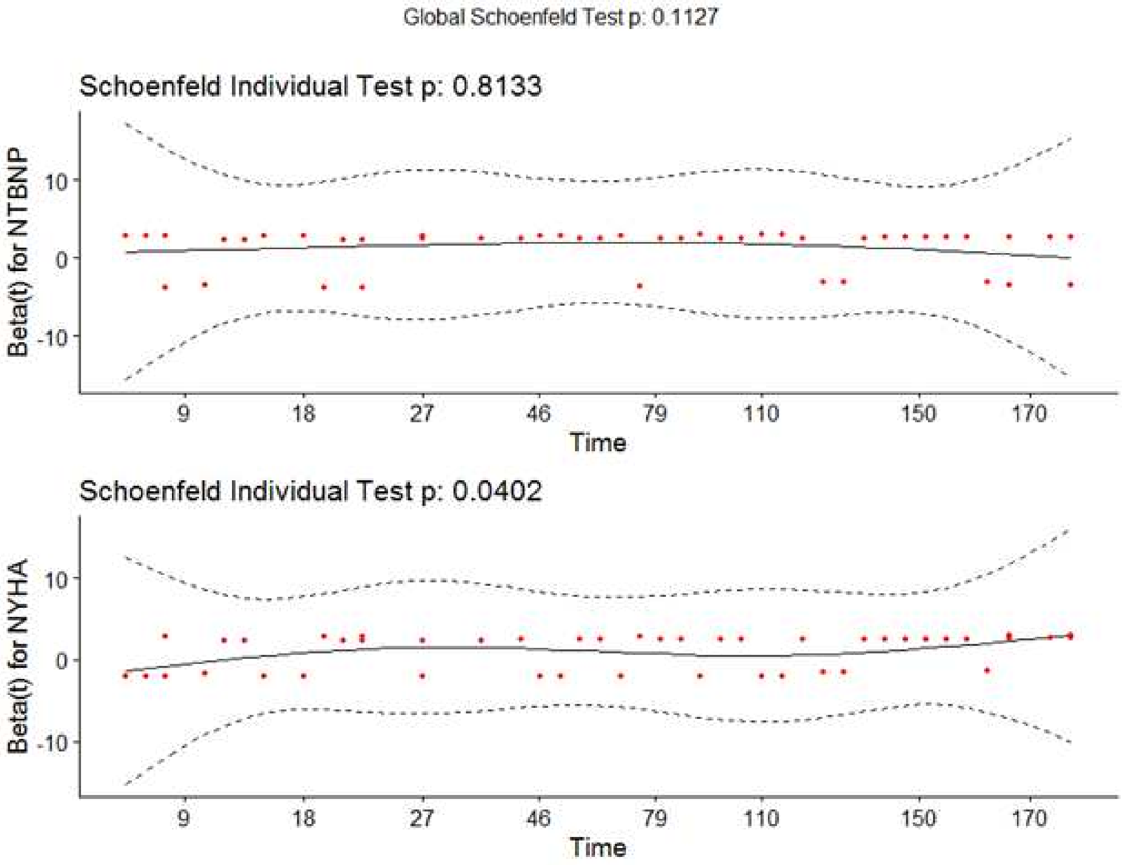


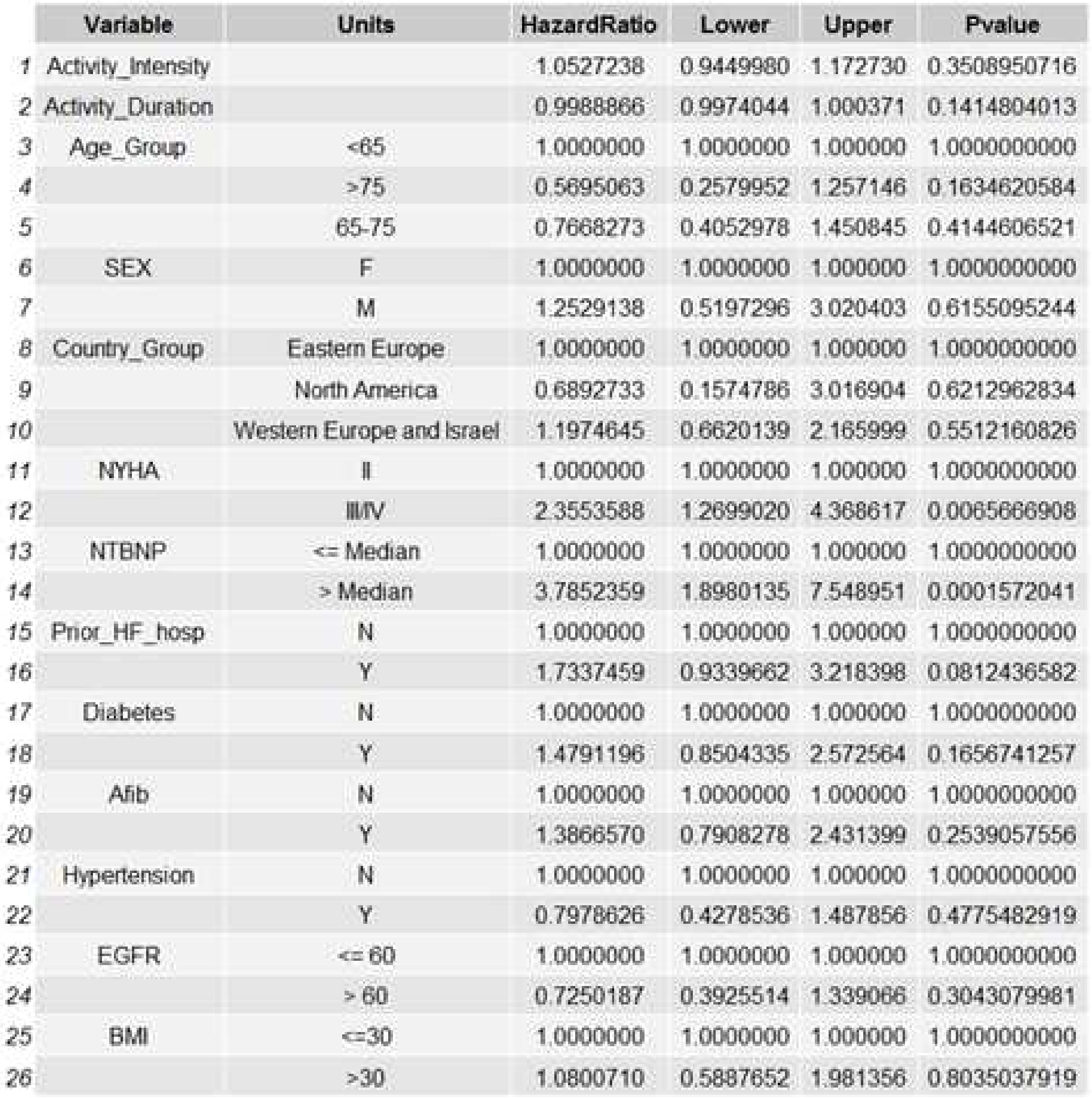
Figure 5.1: Schoenfeld residuals for Cox Proportional hazards model for time to first heart failure hospitalization

The plot provides a depiction of the variance-weighted transformation with respect to the Schoenfeld residuals for each of the covariates in the final reduced model, reflecting the scaled and smoothed Schoenfeld residuals. This plot provides an estimate of the coefficient regression over time for each individual covariate. The above plot indicates a significant individual Schoenfeld test for the covariate NYHA (p = 0.0402), however, the overall Global Schoenfeld test is non-significant (p = 0.1127), indicating that we do not have any issues with non-proportionality. Overall, visual inspection of the plots indicate that the plots are reasonably flat, further suggesting that the PH assumption holds and the Cox PH model is appropriate for this data, allowing for the interpretation of the findings of the model.

5.4 Results for the composite of time to first heart failure hospitalization or urgent heart failure visit based on Cox PH model

A Cox PH model was fitted to the data with the variables activity intensity, activity duration, sex, country grouping, NYHA class at baseline, NT-proBNP at baseline, prior heart failure hospitalization, diabetes, atrial fibrillation, hypertension, EGFR, and BMI as independent variables and the time-to-event composite clinical outcome of time to first heart failure hospitalization and urgent heart failure visit as the dependent variable which yielded the following results:

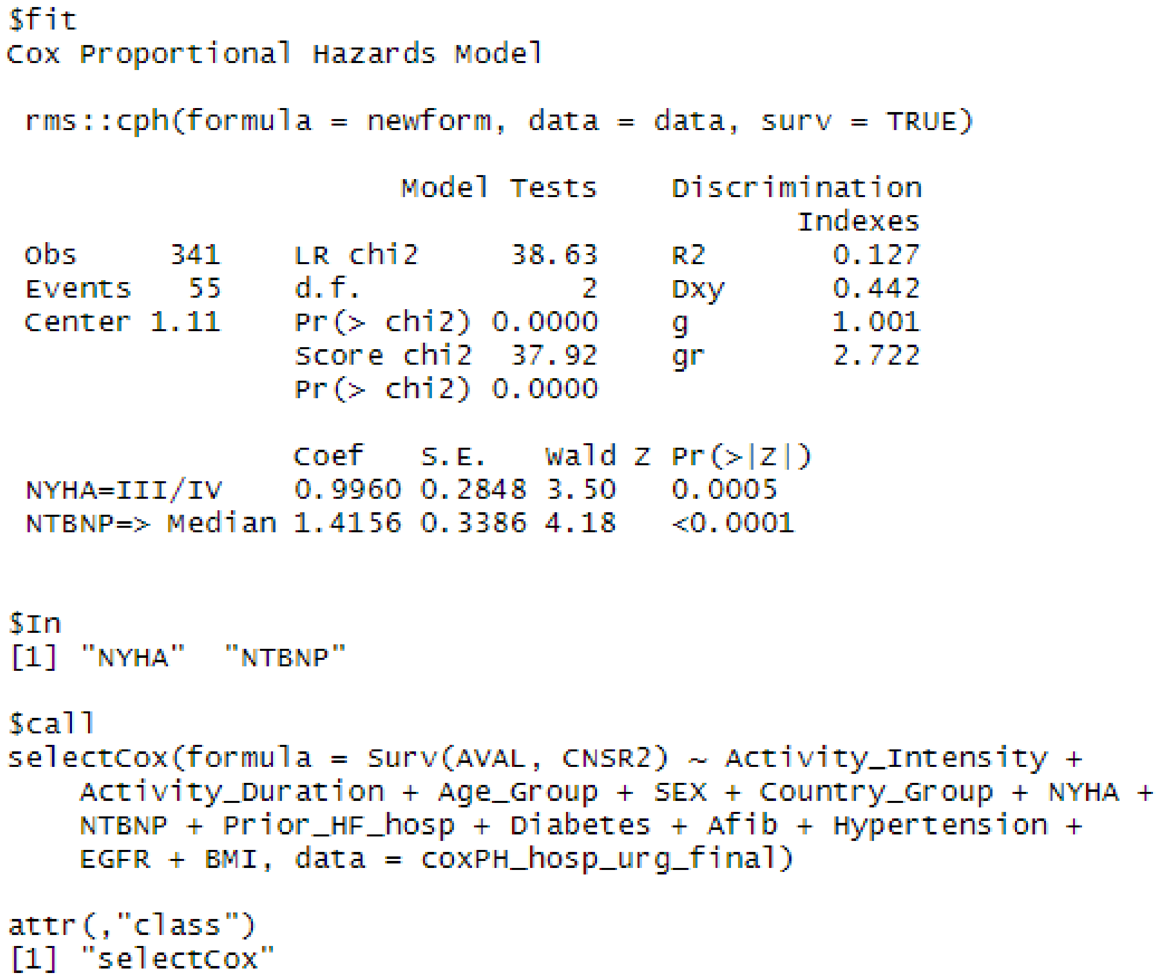
Table 5.4: Cox Proportional Hazards model output for the composite of time to first heart failure hospitalization or urgent heart failure visit



	Variable	Units	HazardRatio	Lower	Upper	Pvalue
1	Activity_Intensity		1.0527238	0.9449980	1.172730	0.3508950716
2	Activity_Duration		0.9988866	0.9974044	1.000371	0.1414804013
3	Age_Group	<65	1.0000000	1.0000000	1.000000	1.0000000000
4		>75	0.5695063	0.2579952	1.257146	0.1634620584
5		65-75	0.7668273	0.4052978	1.450845	0.4144606521
6	SEX	F	1.0000000	1.0000000	1.000000	1.0000000000
7		M	1.2529138	0.5197296	3.020403	0.6155095244
8	Country_Group	Eastern Europe	1.0000000	1.0000000	1.000000	1.0000000000
9		North America	0.6892733	0.1574786	3.016904	0.6212962834
10		Western Europe and Israel	1.1974645	0.6620139	2.165999	0.5512160826
11	NYHA	II	1.0000000	1.0000000	1.000000	1.0000000000
12		III/IV	2.3553588	1.2699020	4.368617	0.0065666908
13	NTBNP	<= Median	1.0000000	1.0000000	1.000000	1.0000000000
14		> Median	3.7852359	1.8980135	7.548951	0.0001572041
15	Prior_HF_hosp	N	1.0000000	1.0000000	1.000000	1.0000000000
16		Y	1.7337459	0.9339662	3.218398	0.0812436582
17	Diabetes	N	1.0000000	1.0000000	1.000000	1.0000000000
18		Y	1.4791196	0.8504335	2.572564	0.1656741257
19	Afib	N	1.0000000	1.0000000	1.000000	1.0000000000
20		Y	1.3866570	0.7908278	2.431399	0.2539057556
21	Hypertension	N	1.0000000	1.0000000	1.000000	1.0000000000
22		Y	0.7978626	0.4278536	1.487856	0.4775482919
23	EGFR	<= 60	1.0000000	1.0000000	1.000000	1.0000000000
24		> 60	0.7250187	0.3925514	1.339066	0.3043079981
25	BMI	<=30	1.0000000	1.0000000	1.000000	1.0000000000
26		>30	1.0800710	0.5887652	1.981356	0.8035037919

The above output yielded NYHA class and NT-proBNP as the most important variables with respect to the clinical outcome of time to first heart failure hospitalization, as indicated by both variables having a p-value ≤ 0.05. The variable prior heart failure hospitalization is trending towards significance, with a p-value ≤ 0.10. A backwards selection approach to the Cox PH model was also fitted to the data yielding the following results:

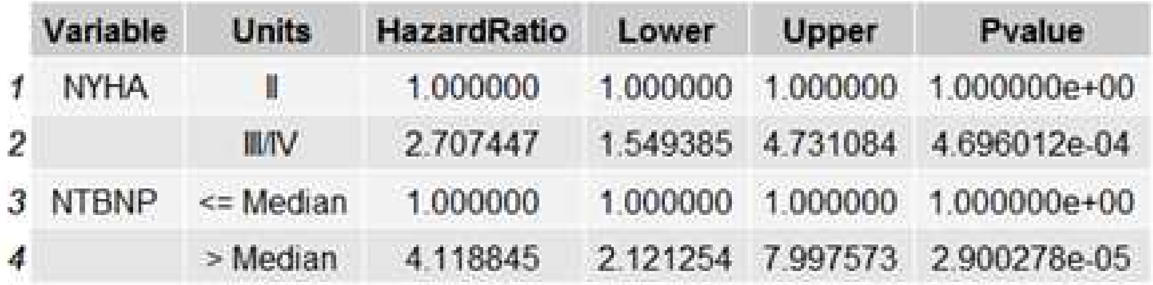
Table 5.5: Cox Proportional Hazards model with backwards selection output for the composite of time to first heart failure hospitalization or urgent heart failure visit



```
$fit
Cox Proportional Hazards Model

 rms::cph(formula = newform, data = data, surv = TRUE)

                    Model Tests     Discrimination
                                           Indexes
 Obs       341    LR chi2      38.63    R2       0.127
 Events     55    d.f.             2    Dxy      0.442
 Center 1.11      Pr(> chi2) 0.0000     g        1.001
                  Score chi2   37.92    gr       2.722
                  Pr(> chi2) 0.0000

                  Coef    S.E.    Wald Z Pr(>|Z|)
 NYHA=III/IV      0.9960 0.2848 3.50    0.0005
 NTBNP=> Median 1.4156 0.3386 4.18    <0.0001


$In
[1] "NYHA"   "NTBNP"

$call
selectCox(formula = Surv(AVAL, CNSR2) ~ Activity_Intensity +
    Activity_Duration + Age_Group + SEX + Country_Group + NYHA +
    NTBNP + Prior_HF_hosp + Diabetes + Afib + Hypertension +
    EGFR + BMI, data = coxPH_hosp_urg_final)

attr(,"class")
[1] "selectCox"
```

The backward selection approach further confirmed NYHA and NT-proBNP as the only significant variables with to the time-to-event composite clinical outcome of time to first heart failure hospitalization and urgent heart failure visit. The Cox PH model was rerun to just include the variables NYHA and NT-proBNP to assess the model features as shown in the following table:

Table 5.6: Cox Proportional Hazards model final reduced output for the composite of time to first heart failure hospitalization or urgent heart failure visit



	Variable	Units	HazardRatio	Lower	Upper	Pvalue
1	NYHA	II	1.000000	1.000000	1.000000	1.000000e+00
2		III/IV	2.707447	1.549385	4.731084	4.696012e-04
3	NTBNP	<= Median	1.000000	1.000000	1.000000	1.000000e+00
4		> Median	4.118845	2.121254	7.997573	2.900278e-05

Both variables remained significant in the final reduced model suggesting these are the appropriate predictors to include. To further assess the appropriateness of the Cox PH model with this data, an assessment of the underlying assumptions of the Cox PH model were performed by examining the Schoenfeld residuals as shown in the figure below [43]:

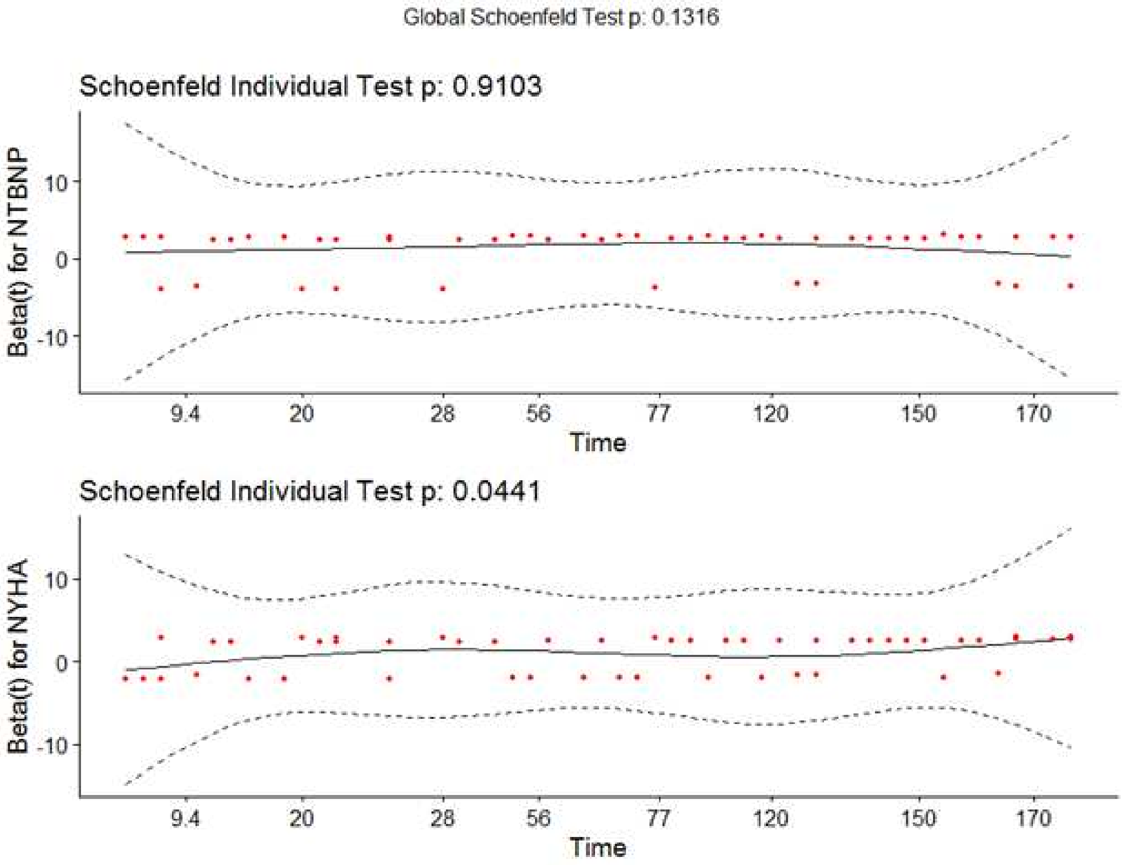


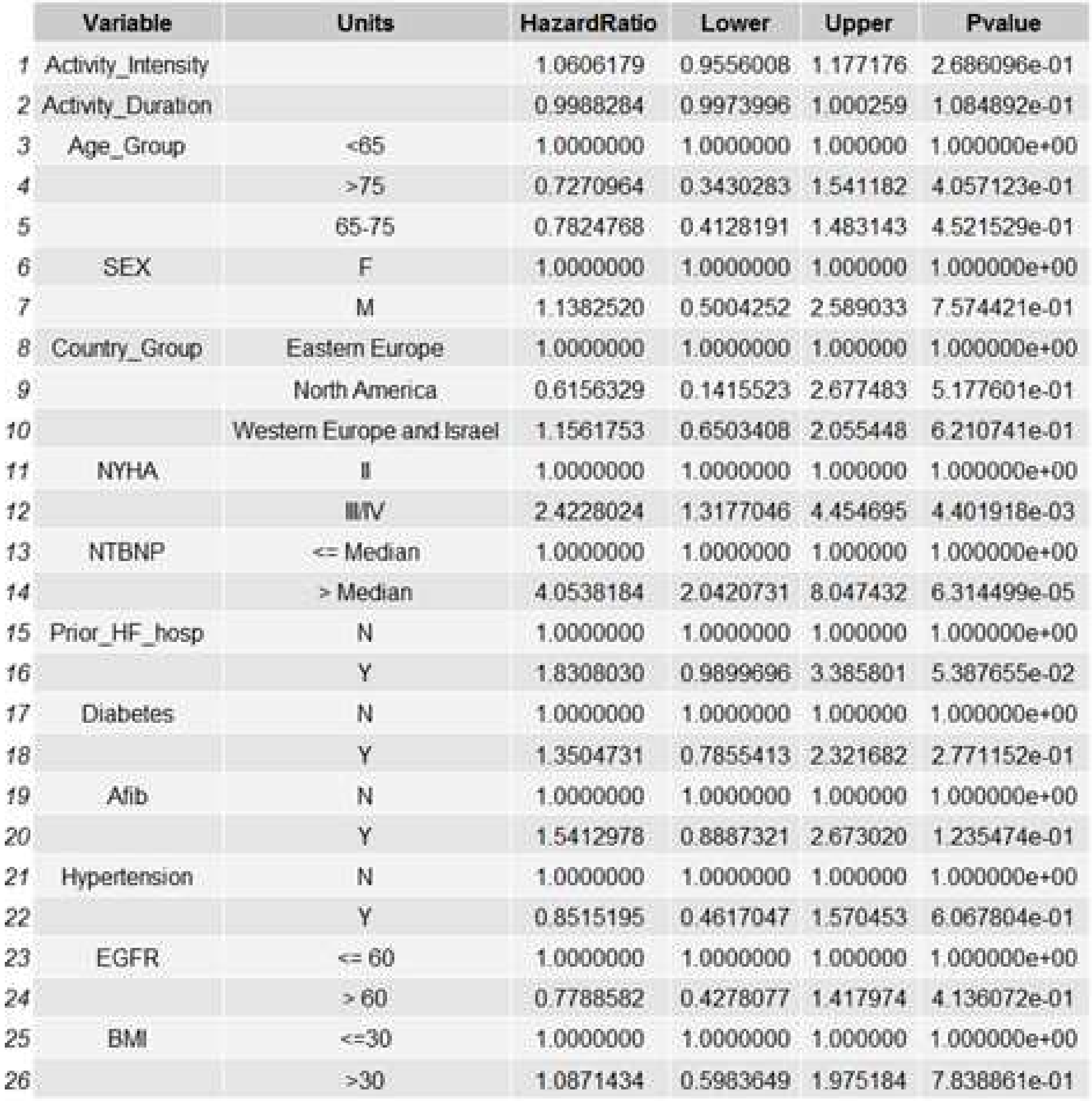
Figure 5.2: Schoenfeld residuals for Cox Proportional hazards model for the composite of time to first heart failure hospitalization or urgent heart failure visit

The plot provides a depiction of the variance-weighted transformation with respect to the Schoenfeld residuals for each of the covariates in the final reduced model, reflecting the scaled and smoothed Schoenfeld residuals. This plot provides an estimate of the coefficient regression over time for each individual covariate. The above plot indicates a significant individual Schoenfeld test for the covariate NYHA (p = 0.0441), however, the overall Global Schoenfeld test is non-significant (p = 0.1316), indicating that we do not have any issues with non-proportionality. Overall, visual inspection of the plots indicate that the plots are reasonably flat, further suggesting that the PH assumption holds and the Cox PH model is appropriate for this data, allowing for the interpretation of the findings of the model.

5.5 Results for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model

A Cox PH model was fitted to the data with the variables activity intensity, activity duration, sex, country grouping, NYHA class at baseline, NT-proBNP at baseline, prior heart failure hospitalization, diabetes, atrial fibrillation, hypertension, EGFR, and BMI as independent variables and the composite of time to first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death as the dependent variable which yielded the following results:

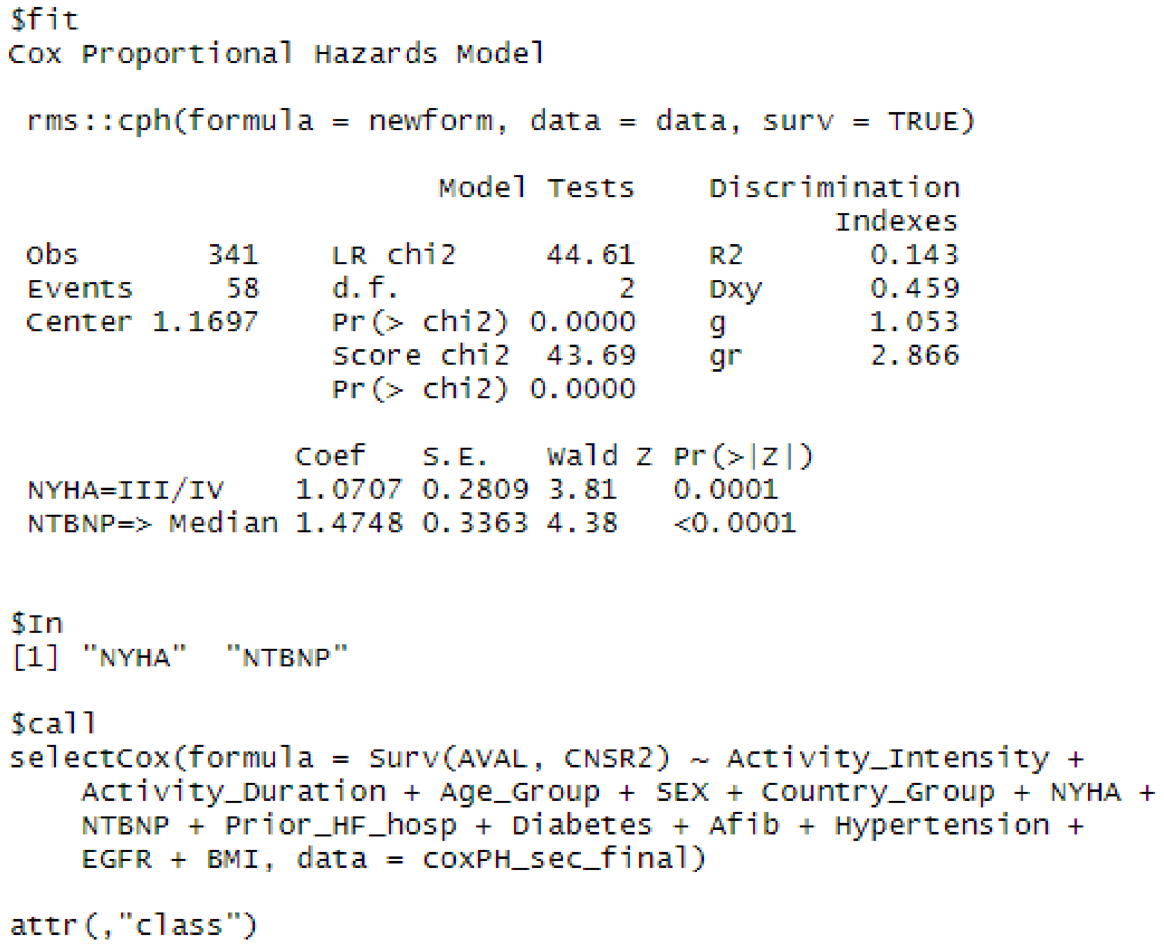
Table 5.7: Cox Proportional Hazards model output for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model



	Variable	Units	HazardRatio	Lower	Upper	Pvalue
1	Activity_Intensity		1.0606179	0.9556008	1.177176	2.686096e-01
2	Activity_Duration		0.9988284	0.9973996	1.000259	1.084892e-01
3	Age_Group	<65	1.0000000	1.0000000	1.000000	1.000000e+00
4		>75	0.7270964	0.3430283	1.541182	4.057123e-01
5		65-75	0.7824768	0.4128191	1.483143	4.521529e-01
6	SEX	F	1.0000000	1.0000000	1.000000	1.000000e+00
7		M	1.1382520	0.5004252	2.589033	7.574421e-01
8	Country_Group	Eastern Europe	1.0000000	1.0000000	1.000000	1.000000e+00
9		North America	0.6156329	0.1415523	2.677483	5.177601e-01
10		Western Europe and Israel	1.1561753	0.6503408	2.055448	6.210741e-01
11	NYHA	II	1.0000000	1.0000000	1.000000	1.000000e+00
12		III/IV	2.4228024	1.3177046	4.454695	4.401918e-03
13	NTBNP	<= Median	1.0000000	1.0000000	1.000000	1.000000e+00
14		> Median	4.0538184	2.0420731	8.047432	6.314499e-05
15	Prior_HF_hosp	N	1.0000000	1.0000000	1.000000	1.000000e+00
16		Y	1.8308030	0.9899696	3.385801	5.387655e-02
17	Diabetes	N	1.0000000	1.0000000	1.000000	1.000000e+00
18		Y	1.3504731	0.7855413	2.321682	2.771152e-01
19	Afib	N	1.0000000	1.0000000	1.000000	1.000000e+00
20		Y	1.5412978	0.8887321	2.673020	1.235474e-01
21	Hypertension	N	1.0000000	1.0000000	1.000000	1.000000e+00
22		Y	0.8515195	0.4617047	1.570453	6.067804e-01
23	EGFR	<= 60	1.0000000	1.0000000	1.000000	1.000000e+00
24		> 60	0.7788582	0.4278077	1.417974	4.136072e-01
25	BMI	<=30	1.0000000	1.0000000	1.000000	1.000000e+00
26		>30	1.0871434	0.5983649	1.975184	7.838861e-01

The above output yielded NYHA class and NT-proBNP as the most important variables with respect to the clinical outcome of time to first heart failure hospitalization, as indicated by both variables having a p-value ≤ 0.05. The variable prior heart failure hospitalization is trending towards significance, with a p-value = 0.0538. A backwards selection approach to the Cox PH model was also fitted to the data yielding the following results:

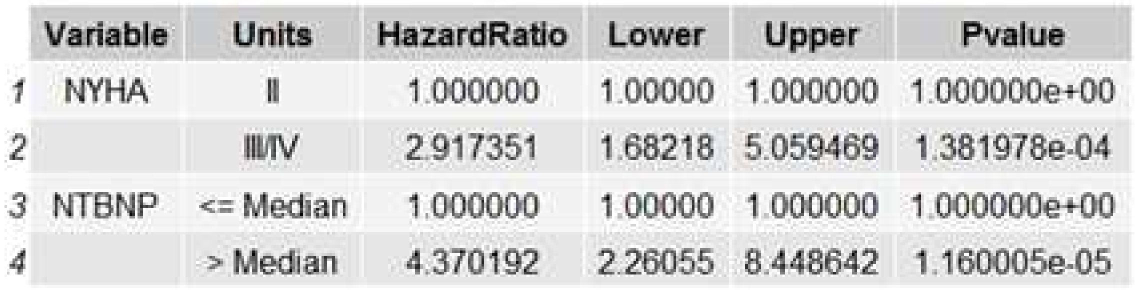
Table 5.8: Cox Proportional Hazards model output for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model



```
$fit
Cox Proportional Hazards Model

 rms::cph(formula = newform, data = data, surv = TRUE)

                          Model Tests    Discrimination
                                                Indexes
 Obs        341     LR chi2     44.61    R2        0.143
 Events      58     d.f.            2    Dxy       0.459
 Center 1.1697      Pr(> chi2) 0.0000    g         1.053
                    Score chi2  43.69    gr        2.866
                    Pr(> chi2) 0.0000

                 Coef   S.E.   Wald Z Pr(>|Z|)
 NYHA=III/IV     1.0707 0.2809 3.81   0.0001
 NTBNP=> Median 1.4748 0.3363 4.38   <0.0001


$In
[1] "NYHA"  "NTBNP"

$call
selectCox(formula = Surv(AVAL, CNSR2) ~ Activity_Intensity +
    Activity_Duration + Age_Group + SEX + Country_Group + NYHA +
    NTBNP + Prior_HF_hosp + Diabetes + Afib + Hypertension +
    EGFR + BMI, data = coxPH_sec_final)

attr(,"class")
```

The backward selection approach further confirmed NYHA and NT-proBNP as the only significant variables with to the time-to-event composite clinical outcome of time to first heart failure hospitalization and urgent heart failure visit. The Cox PH model was rerun to just include the variables NYHA and NT-proBNP to assess the model features as shown in the following table:

Table 5.9: Cox Proportional Hazards model final reduced output for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model



	Variable	Units	HazardRatio	Lower	Upper	Pvalue
1	NYHA	II	1.000000	1.00000	1.000000	1.000000e+00
2		III/IV	2.917351	1.68218	5.059469	1.381978e-04
3	NTBNP	<= Median	1.000000	1.00000	1.000000	1.000000e+00
4		> Median	4.370192	2.26055	8.448642	1.160005e-05

Both variables remained significant in the final reduced model suggesting these are the appropriate predictors to include. To further assess the appropriateness of the Cox PH model with this data, an assessment of the underlying assumptions of the Cox PH model were performed by examining the Schoenfeld residuals as shown in the figure below [43]:

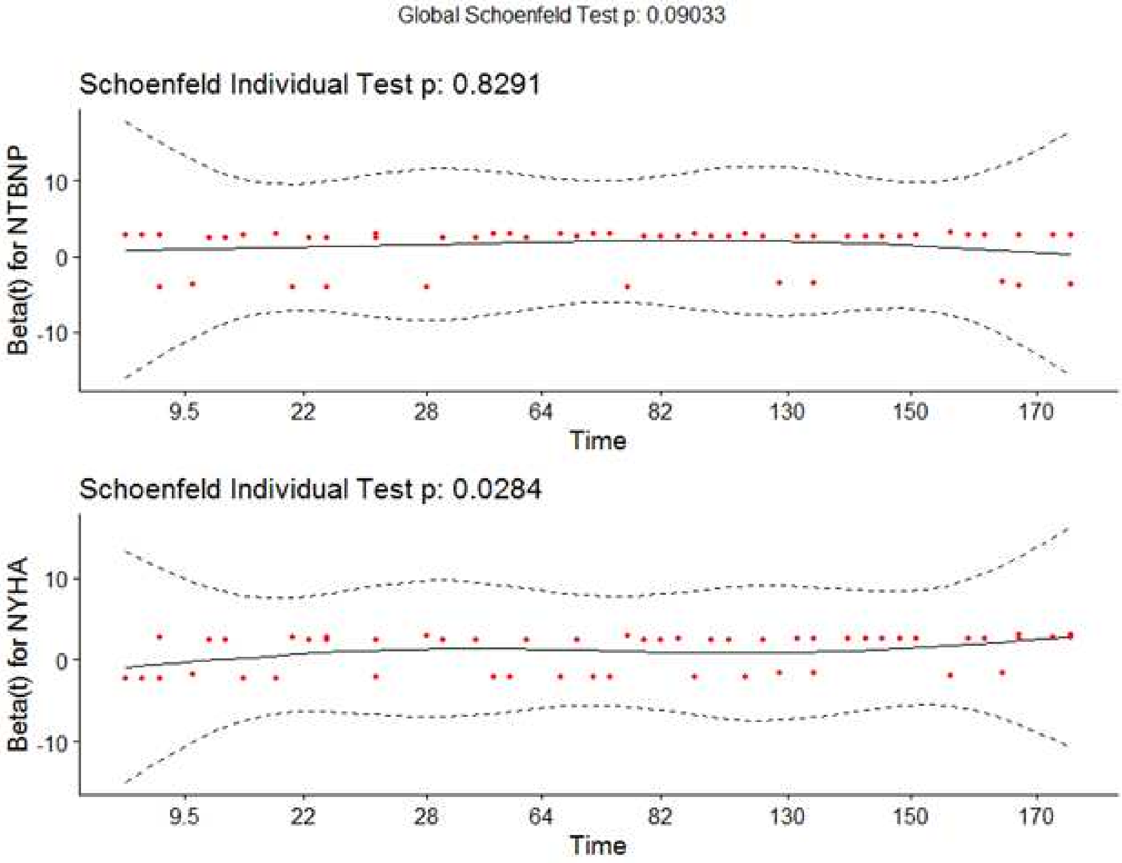


Figure 5.3: Schoenfeld residuals for Cox Proportional hazards model for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Cox PH model

The plot provides a depiction of the variance-weighted transformation with respect to the Schoenfeld residuals for each of the covariates in the final reduced model, reflecting the scaled and smoothed Schoenfeld residuals. This plot provides an estimate of the coefficient regression over time for each individual covariate. The above plot indicates a significant individual Schoenfeld test for the covariate NYHA (p = 0.0284), however, the overall Global Schoenfeld test is non-significant (p = 0.09), indicating that we do not have any issues with non-proportionality. Overall, visual inspection of the plots indicate that the plots are reasonably flat, further suggesting that the PH assumption holds and the Cox PH model is appropriate for this data, allowing for the interpretation of the findings of the model.

5.6 Discussion based on Cox PH model

Across all of the Cox PH models for the three dependent variable clinical outcomes, NT-proBNP and NYHA class appeared as the two significant predictor variables. This is not surprising given the well documented and strong predictive value of these variables with respect to clinical outcomes of heart failure [44]. Furthermore, both of these variables convey information as to the state of severity of a heart failure patient and it is known that they are positively correlated [45] and thus one could expect for them to appear implicated in the prediction of clinically relevant outcomes data.

The variables diabetes and heart failure hospitalization were identified as trending towards significance in the analyses, suggesting that they may play an important role in the prediction of clinical outcomes in heart failure. The relationship between diabetes and heart failure is further documented in the literature, with a particular focus on heart failure patients with diabetes having higher rates of clinical outcomes as compared to their heart failure without diabetes counterparts [27]. Additionally, the impact of a previous history of heart failure on further predicting follow-up heart failure is well documented given that once a patient suffers from a heart failure event and needs to be hospitalization, this initial heart failure hospitalization is associated with a further increase in additional heart failure clinical outcomes [26].

While activity duration and activity intensity were not identified as significant variables for the prediction of these clinical outcomes, it is possible we do not have the appropriate power to detect their effects as part of the Cox PH model. Both NT-proBNP and NYHA class have been well documented in the literature with regards to their strong association to clinical outcomes in heart failure [44] and it is possible that activity intensity and activity duration, as identified by the wearable devices, do in fact provide additional useful complimentary information towards the prediction of clinical outcome events, however, this association to clinical outcome events may not be as strong as with NT-proBNP and NYHA class. In particular, as seen in the results for activity intensity, there is a wide confidence interval and thus it is possible that we need further patients to adequately assess the effect. Additionally, this ranking of the importance of activity intensity and activity duration can be further assessed utilizing the random survival forest algorithm.

Chapter 6

Methods: Random Survival Forest

In an attempt to further understand the survival data, a separate machine learning method, the random survival forest, was applied to the data. In addition to the classification work presented earlier, a separate stream of research examined the utility of wearables and biosensors to improve predictive accuracy of models relating to clinical outcome events. Furthermore, the random survival forest was used to assess the ranking of variable importance, including physical activity duration and intensity, on the prediction of clinical outcome events and to assess whether an improvement can be made upon the earlier presented Cox PH models.

The random forest (RF) is a machine learning and statistical learning algorithm that does not assume an underlying distribution to be specified between the independent variables and the dependent variable. It functions as an ensemble of unique tree-based learners, aggregating individual decision trees together and averaging their operating characteristics. Through this iterative process, the RF can identify key patterns and interactions amongst variables while reducing variance due to the aggregation of the individual decision trees versus simply taking the results from one individual tree. In addition, this aggregation of individual trees reduces the likelihood that an individual tree will simply overfit to the data. In the case of survival data, the random survival forest (RSF) algorithm extends upon the random forest methodology to take into account the specifics of time-to-event survival analysis data [46, 47]. Furthermore, in comparison to the semi-parametric Cox proportional hazards model, the RSF is non-parametric in nature and can be particularly useful if data is non-proportional or if the analysis should be conducted without specifying a certain distribution beforehand.

In an attempt to help validate the model, reduce overfitting, and improve variance estimates, the RSF can utilize bootstrap aggregation, in which iterative training samples are created through sampling with replacement in order for the model to learn from and create decision trees [47]. Out-of-bag (OOB) error is then calculated from the bootstrapped aggregation for each training sample and then averaged across the samples to calculate the out-of-bag estimate of the overall predictive performance of the model [48]. The following steps are part of the random survival forest: 1) A bootstrapped sample of data is created from the original dataset and the pre-specified number of trees identified a-priori, 2) A decision tree is grown and subsequently within each tree and for each node, a number of predictors is used to assess the splitting. Given the survival nature of data here, a log-rank splitting rule is used, 3) The splits continue within the tree until the final node of the tree is converged upon, and 4) The results are combined and averaged together to compute the final ensemble cumulative function for all patients [47]. With respect to the outcomes of the RSF, there are two main variable importance measures to rank the importance of predictive variables: the permutation variable importance (VIMP) and minimal depth. The permutation VIMP is derived in the following manner: 1) Initially calculate the prediction error for a tree utilizing patients who were not included in the subsample of that specific tree, 2) Create a permutation of the values for a specific variable for those observation, 3) Re-calculate the prediction error of the tree utilizing the values of the now permuted variable and compute differences between predictive accuracy, and 4) This process is repeated across variables and trees and the results combined and averaged together [47]. The results of the VIMP yield values for each variable that can be grouped as either ≤ 0 (negative or zero values) or > 0 (positive values). A variable yielding an overall positive VIMP is indicative that the variable is contributing positive predictive importance. A variable yielding an overall negative or zero-value VIMP is indicative that the variable is not contributing positive predictive importance for the model. It is often suggested to remove variables that yield a VIMP ≤ 0 from the model and to re-run the model in an attempt to improve the predictive accuracy of the overall model and include only variables with a positive VIMP > 0 in the final model. This then yields the final reduced model [49]. The minimal depth approach assesses the overall predictive utility of a variable by calculating how far down the node the variable is as compared to the root node of the tree. The closer the split upon the variable is to the root node, the smaller the minimal depth and the more predictive the effects of the variable is on the outcome dependent variable [50]. Furthermore as outputs to the RSF analysis, the OOB prediction error is used to provide a quantitative metric as to what extent the outcome can be explained and predicted by the included baseline variables. The output for range for the OOB prediction error is [0% - 50%]. In the case of an OOB prediction error of 50%, the baseline information does not contain any relevant information for the predicting the outcome and essentially that the model performance at predicting the outcome is the same as flipping a coin [48].

The following steps were taken with respect to training the random survival forest algorithm. Given the probabilistic nature of iterative repetitions of the algorithmic process, an initial seed was set for the algorithm prior to subsequent simulations and analyses. Data was randomly split into a testing and training set, with 80% of the data comprising the training set and 20% of the data comprising the testing set. A series of simulations were then conducted to identify the model with the lowest out-of-bag (OOB) prediction error from the bootstrapped aggregation, which provides the lowest mean prediction error for each iteration of the training sample (Gareth et al., 2013). The simulation approaches the task of optimizing the OOB prediction error by running through an iterative succession of hyperparameters to tune the model through a grid search task. More specifically, the parameters of node size, nsplit, and mtry are modulated in the simulation with the following parameters to perform the grid search optimization [51]:

```
nodesize<-c(10,20,35,50,70,85,100,120,150,180,190,200,210,220)
nsplit<-c(2,3,4,5,6,7,8,9,10,15,20)
mtry<-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)
```

The R package randomForestSRC version 3.1.0 was used to perform the random survival forest algorithm [51]. Within the algorithm, node size refers to the average overall forest node size, nsplit refers to the value used to identify random splitting, and mtry identifies the number of the variable subset that is randomly selected during each iteration of the model (in this case with an upper limit of 15 as for this dataset there are a total of 15 independent variables included in the model) [52]. These combinations of these hyperparameters are then iteratively assessed throughout the algorithm with the OOB prediction error stored in a grid. As the final step, the model with the least OOB prediction error is selected and these specific values of the hyperparameters node size, nsplit, and mtry are then used to identify and examine the final model. Across the simulations, the logrank splitting rule is used by the algorithm to decide upon splits of the tree. The value of the hyperparameter ntree was set equal to 500 to avoid high variance in the resulting VIMPs [52]. Furthermore, a minimal depth criterion is used to assess the importance of variables identified by the random survival forest. The minimal depth is then averaged across the trees to provide a reliable measure, with lower values indicative of the most predictive variables for the clinical outcomes of interest. For the purposes of refining the models and selecting the final variables included in the reduced model, the VIMP will primarily be used.

6.1 Results for time to first heart failure hospitalization based on Random Survival Forest

A random survival forest algorithm was fitted to the independent variables of NT-proBNP at baseline, NYHA class, activity duration, activity intensity, prior heart failure hospitalization, atrial fibrillation, diabetes, age group, sex, EGFR at baseline, BMI at baseline, country group, and hypertension at baseline to predict the dependent clinical outcome variable of heart failure hospitalization, yielding the below output for the VIMP:

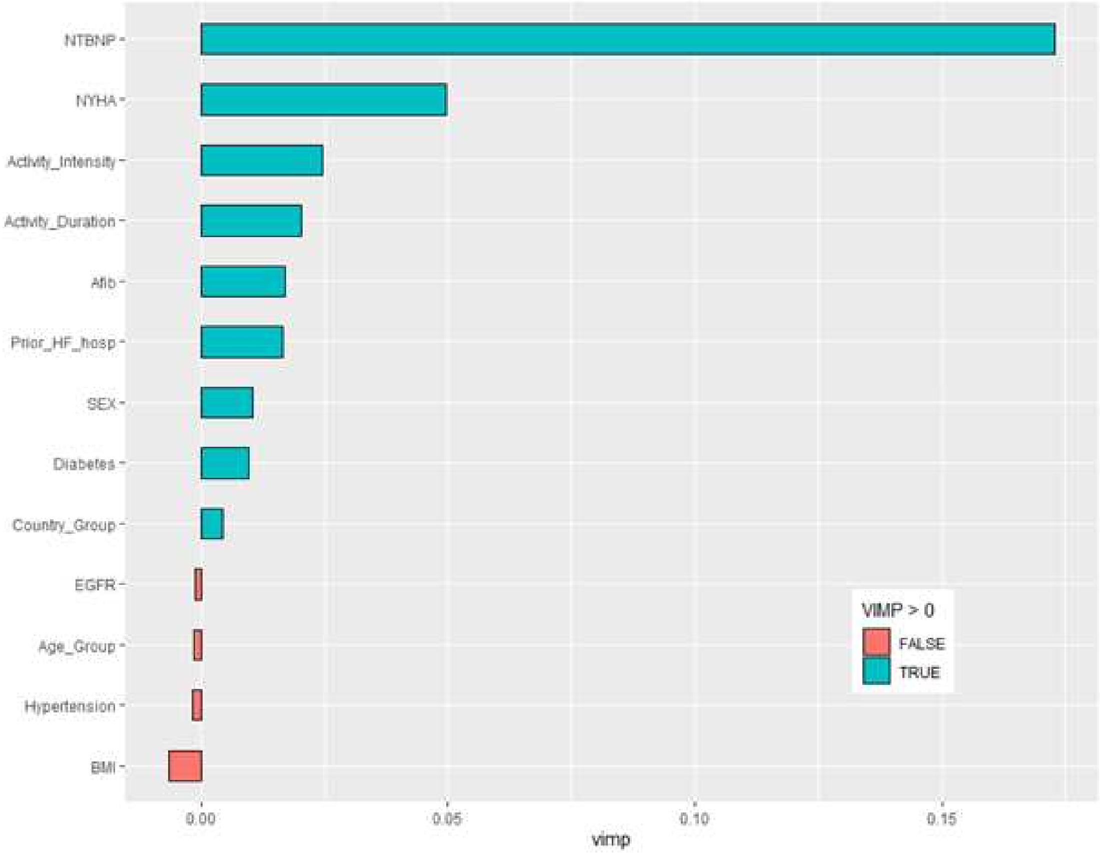
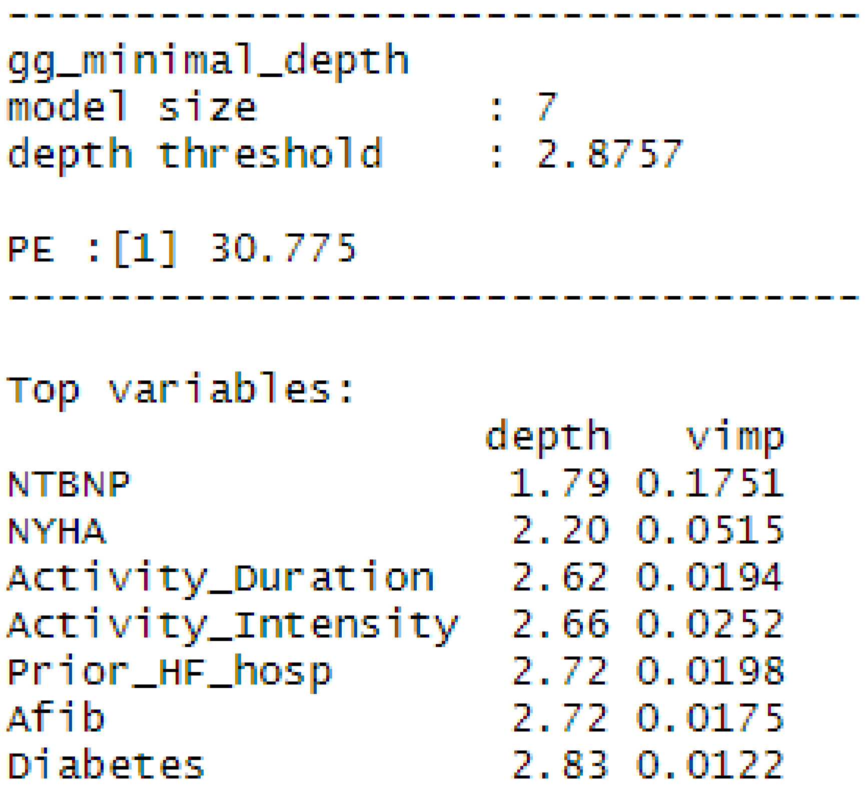


Figure 6.1: Random survival forest VIMP output for clinical outcome of heart failure hospitalization

After following the simulation process grid search denoted above, the model was tuned to the hyperparameters values of nsplit = 2, node size = 20, and mtry = 4, yielding an OOB prediction error for the overall model of 0.3078. The results categorized the variables into two groupings; those with a positive VIMP and those with a negative VIMP. The variables with positive VIMPs for the outcome of time to first heart failure hospitalization included NT-proBNP, NYHA class, activity intensity, prior heart failure hospitalization, activity duration, atrial fibrillation, sex, and history of diabetes. Variables with a negative VIMP include baseline EGFR, age group, hypertension, and BMI. Additionally, a minimal depth approach is used to assess the random survival variable importance through another measure to assess similarity of findings [50]. The output for this approach is seen below in graphical and tabular format:

Table 6.1: Random survival forest minimal depth and VIMP scores for clinical outcome of heart failure



```
----------------------------------
gg_minimal_depth
model size         : 7
depth threshold    : 2.8757

PE :[1] 30.775
----------------------------------

Top variables:
                   depth   vimp
NTBNP               1.79 0.1751
NYHA                2.20 0.0515
Activity_Duration   2.62 0.0194
Activity_Intensity  2.66 0.0252
Prior_HF_hosp       2.72 0.0198
Afib                2.72 0.0175
Diabetes            2.83 0.0122
```

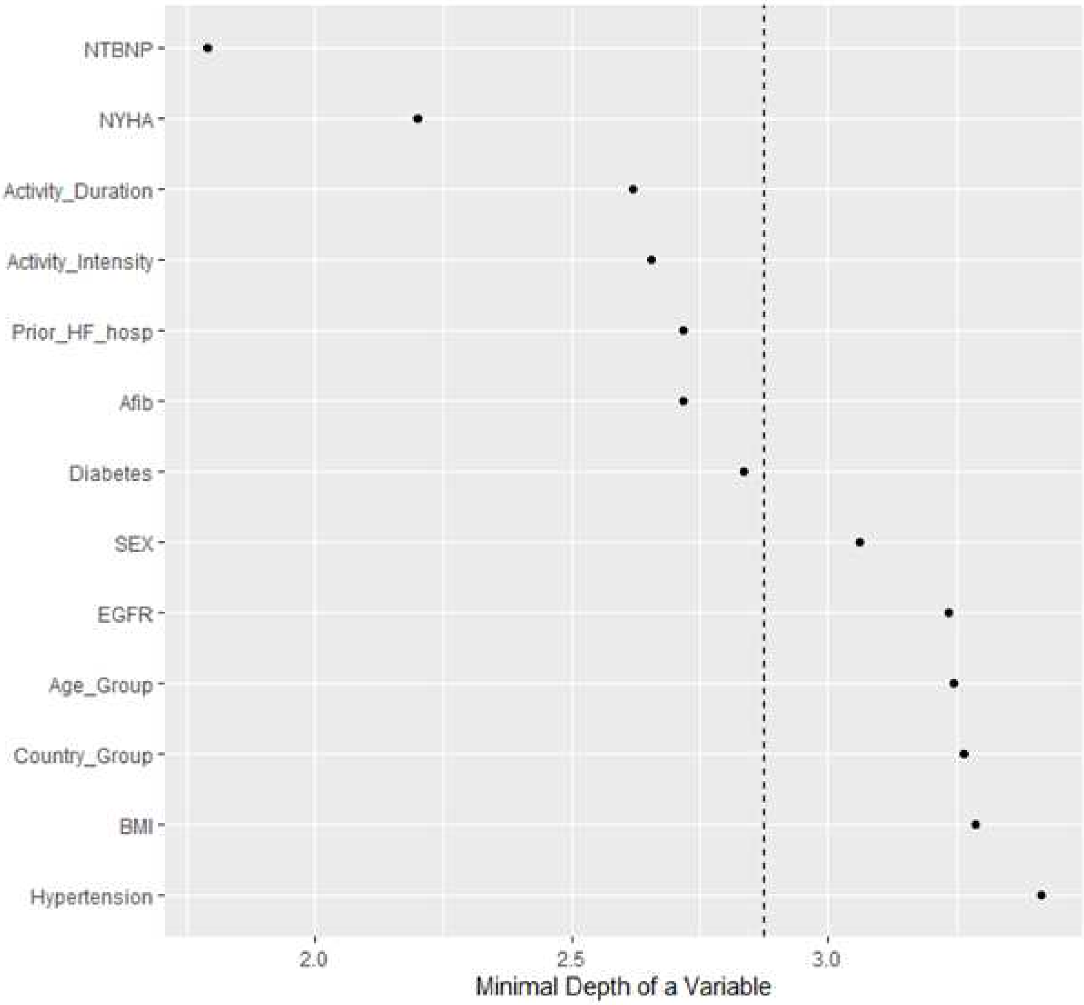


Figure 6.2: Random survival forest minimal depth output for clinical outcome of heart failure hospitalization

The minimal depth approach identifies the same variables as important except sex and baseline EGFR, which are not ranked amongst the most important set of variables. The output above indicates that the same highest rank variables were identified by both the VIMP and the minimal depth, including NT-proBNP, NYHA class, activity duration, and activity intensity. Plotting both the minimal depth and VIMP together allows for a comparison of the selected variables and their ranking between the two variable importance measures:

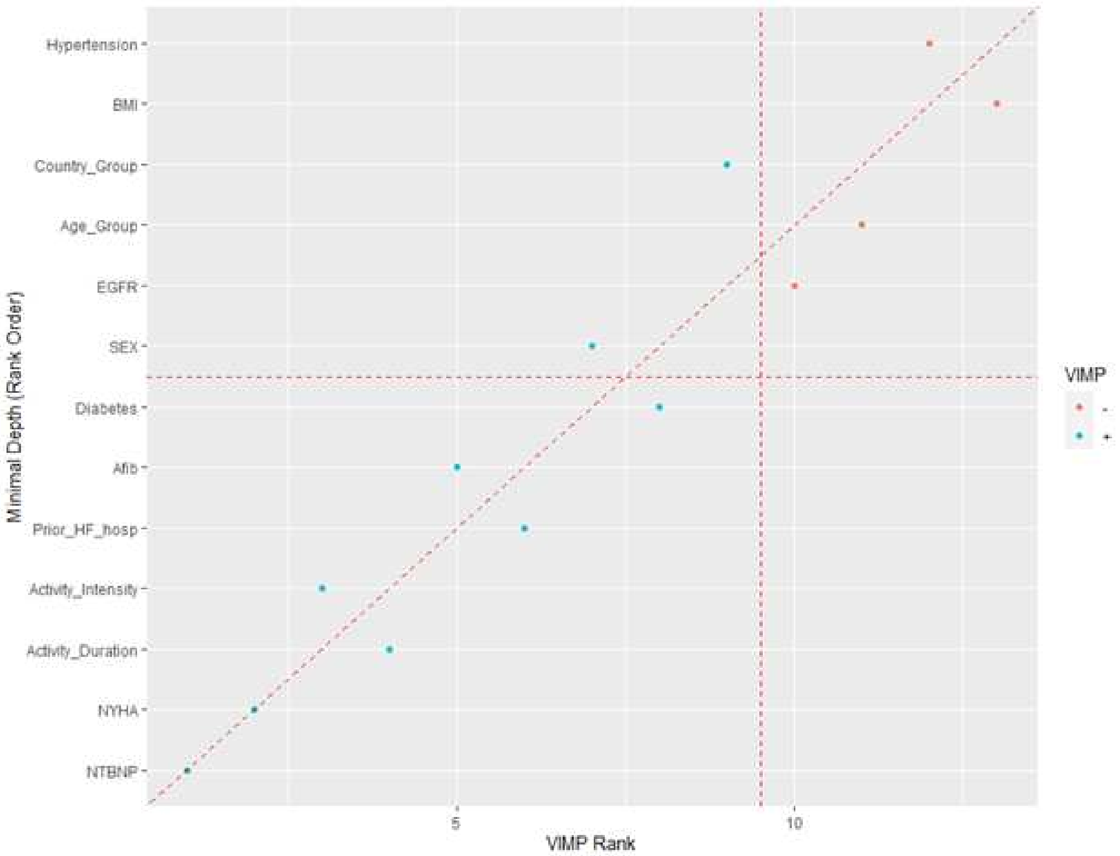


Figure 6.3: Random survival forest minimal depth and VIMP output for clinical outcome of heart failure hospitalization

Concordance between the VIMP and minimal depth is indicated by variables falling on the diagonal line in the plot above, which occurs for most of the variables listed suggesting that the ranking of variables by both variable importance measures is similar. Perfect agreement between the two indices would occur if the variables all fall on the diagonal line. In this case, there is strong overlap between the variables and their importance as measured by the two indices. Given that these variables with VIMP ≤ 0 are not contributing to the predictive power of the model, these variables are removed, and the reduced model is then re-run to assess the appropriateness of including them in the final model [49]. Improvement as quantified by a reduction in OOB error in the reduced model is confirmation that the removal of these variables improves predictive accuracy. The VIMP from the final model with the variables baseline EGFR, age group, BMI, and hypertension removed confirms that all variables left in the reduced model now have a positive VIMP as shown in the following VIMP plot:

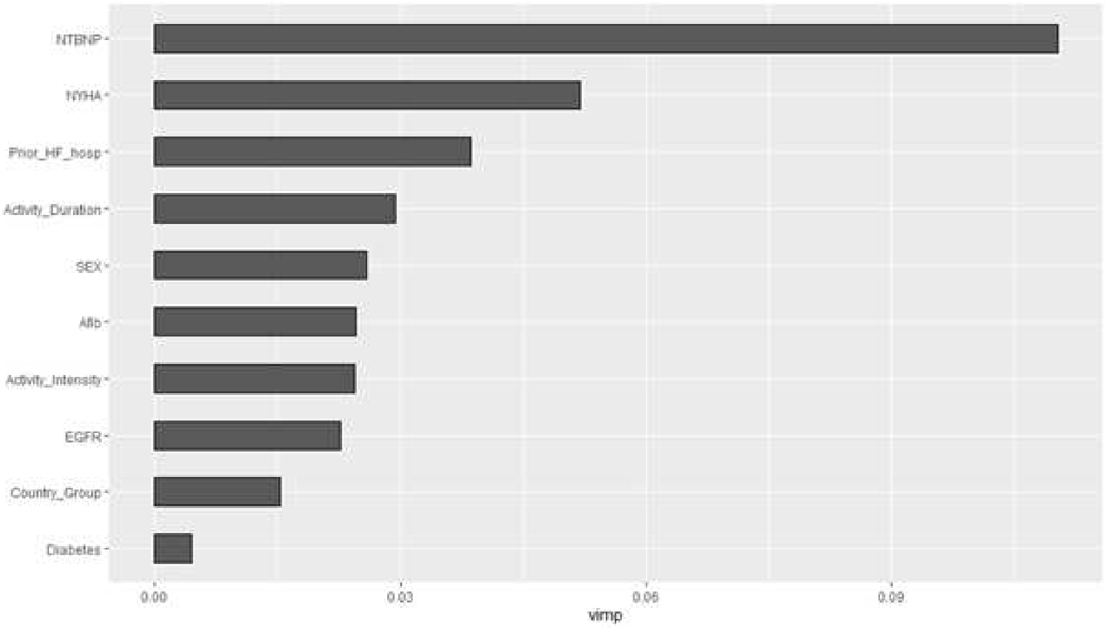


Figure 6.4: Random survival forest final reduced VIMP output for clinical outcome of heart failure hospitalization

As all of the variables have a VIMP > 0, the above VIMP plot indicates that all variables in the model contribute to an improvement in the predictive power of the model. This is also confirmed by an improvement in the OOB predictive error which is 0.2987 for the model. Of importance to note here is the ranking of the variables, with NT-proBNP identified as the most important variable in predicting the clinical outcome of heart failure hospitalization, followed by NYHA class at baseline, prior history of heart failure hospitalization, and then activity duration. Further examination into the partial dependence plots is performed to better understand the relationship between the variables contributing the most improvement in the predictive power of the model and the clinical outcome variable of heart failure hospitalization [53].

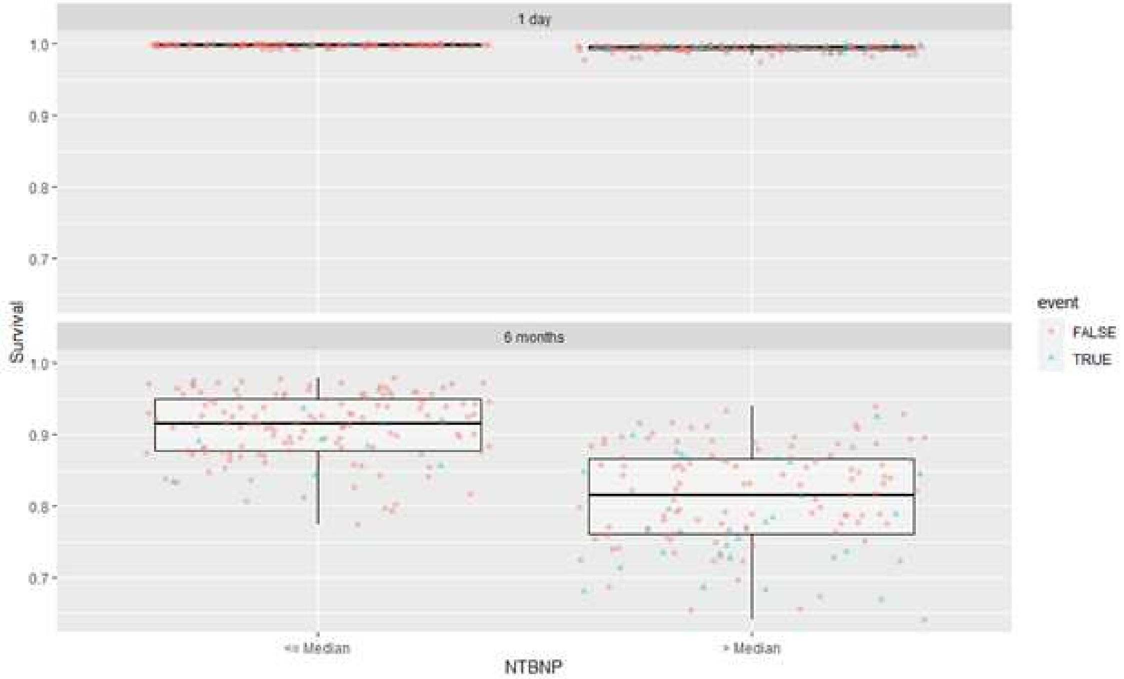


Figure 6.5: Partial dependence plot for baseline NT-proBNP for clinical outcome of heart failure hospitalization

In the above partial dependence plot examining the relationship between NT-proBNP and survival with respect to the clinical outcome of time to first heart failure hospitalization, patients with baseline NT-proBNP less than or equal to the median have higher survival than patients who have baseline NT-proBNP greater than the median as a predictive variable for the clinical outcome of time to heart failure hospitalization. Furthermore, fewer patients in the less than or equal to the median baseline NT-proBNP group had the outcome event (10 out of 173 patients or 5.8%) as compared to those in the greater than median baseline NT-proBNP group (39 out of 174 patients or 22.4%). NYHA class was identified as the second most important variable and the following partial dependence plot examines survival across the variable:

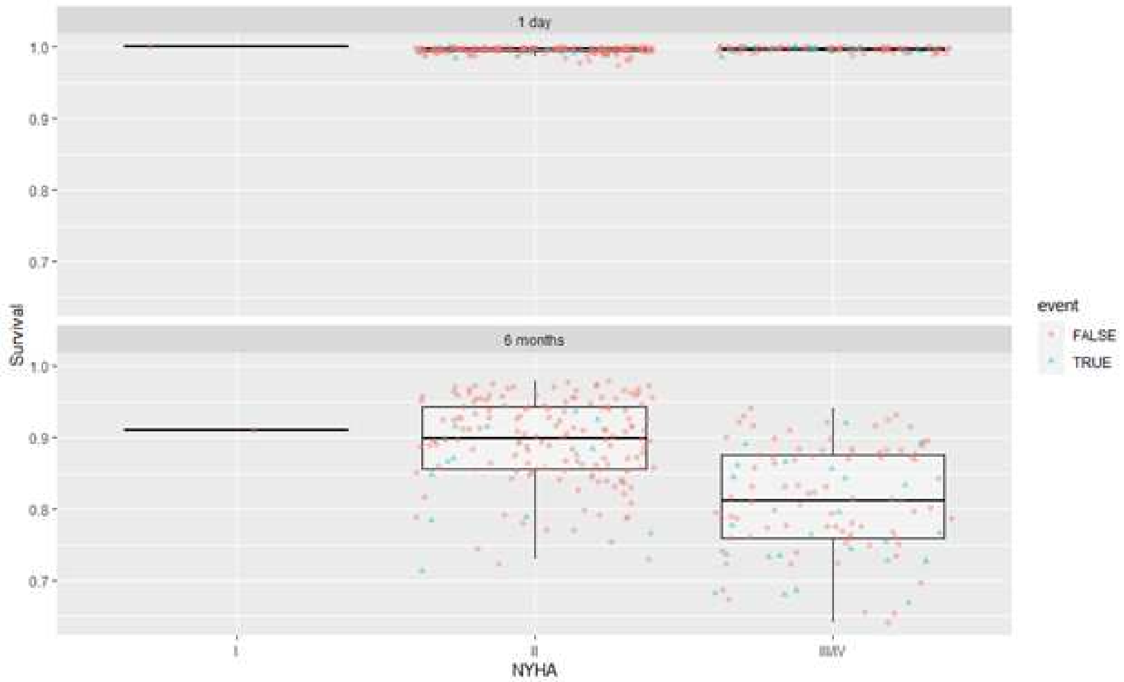


Figure 6.6: Partial dependence plot for baseline NYHA class for clinical outcome of heart failure hospitalization

The grouping of patients across NYHA class showed a trend for which patients who fell into the more severe categories of NYHA classes (Class III/IV) had more events and lower survival than patients who were in NYHA class II. More specifically, there were 17 clinical events in the NYHA class I and II patients out of 208 patients (8.2%). Out of the NYHA class III/IV group, there were 32 patients out of a total of 139 patients who had the clinical event (23.0%). Prior history of heart failure hospitalization was identified as the third most important variable and the following figures examines survival across the variable:

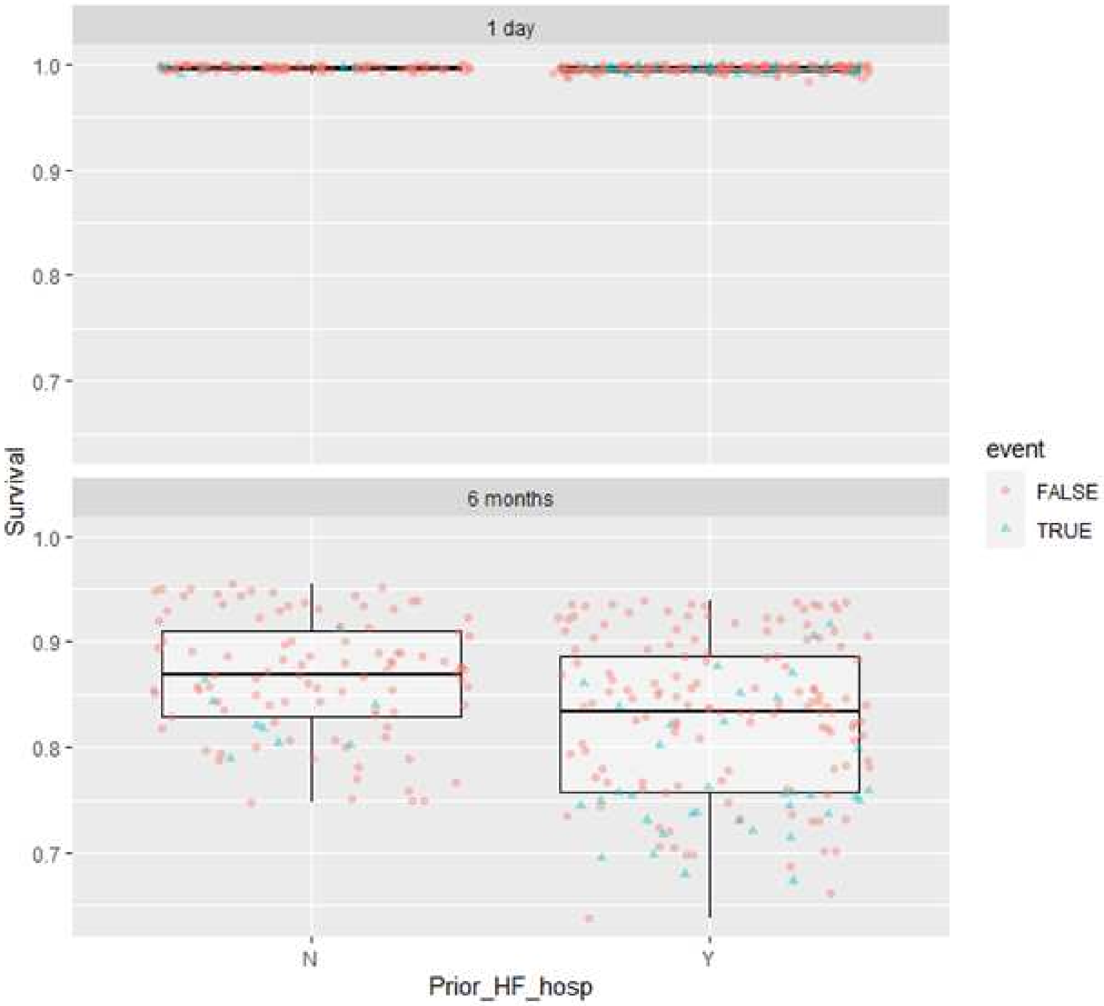


Figure 6.7: Partial dependence plot for prior history of heart failure hospitalization class for clinical outcome of heart failure hospitalization

Patients with no history of prior hospitalization had less events and a higher overall survival (14 clinical outcome events out of a total of 135 patients or 10.4%). Patients with a history of prior hospitalization had more events and a lower overall survival (35 clinical outcome events out of a total of 212 patients or 16.5%). The relationship of activity duration to survival is examined in the following partial dependence plot:

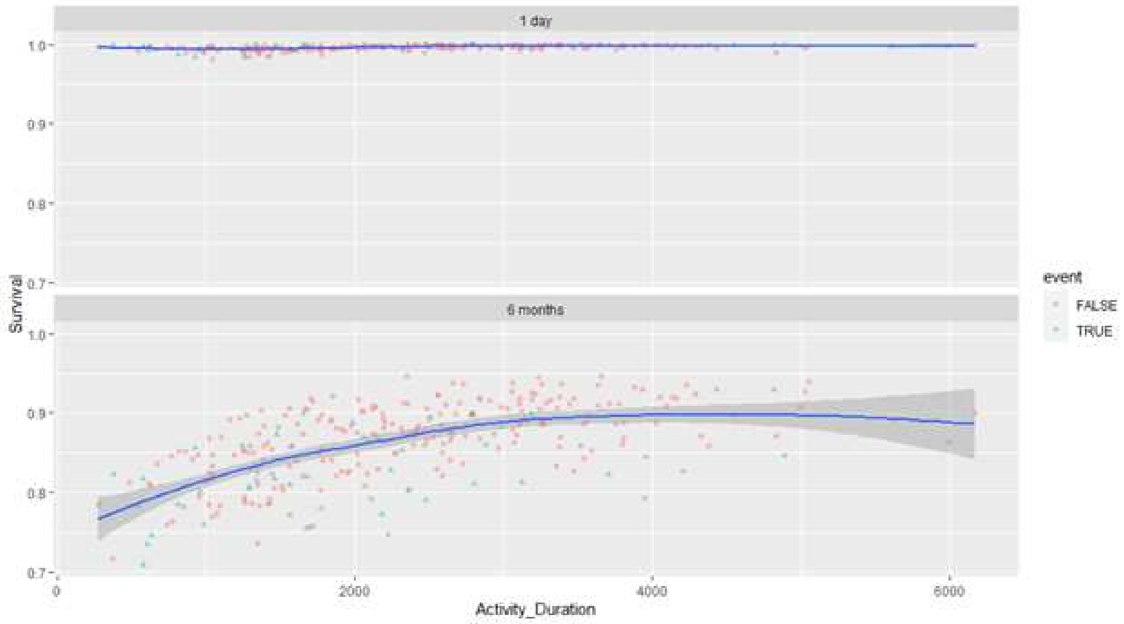


Figure 6.8: Partial dependence plot for activity duration for clinical outcome of heart failure hospitalization

The function of survival by activity duration with respect to the clinical outcome of heart failure hospitalization suggests that patients with lower rates of activity duration, particularly less than 3000 seconds or 50 minutes per day, have lower rates of survival. The function seems to plateau at activity duration rates greater than 3000 seconds or 50 minutes per day, however, it is possible this may be due to less patients at the higher levels of activity duration given that the confidence intervals widen. The relationship of activity intensity to the clinical outcome of time to heart failure hospitalization is examined in the following partial dependence plot:

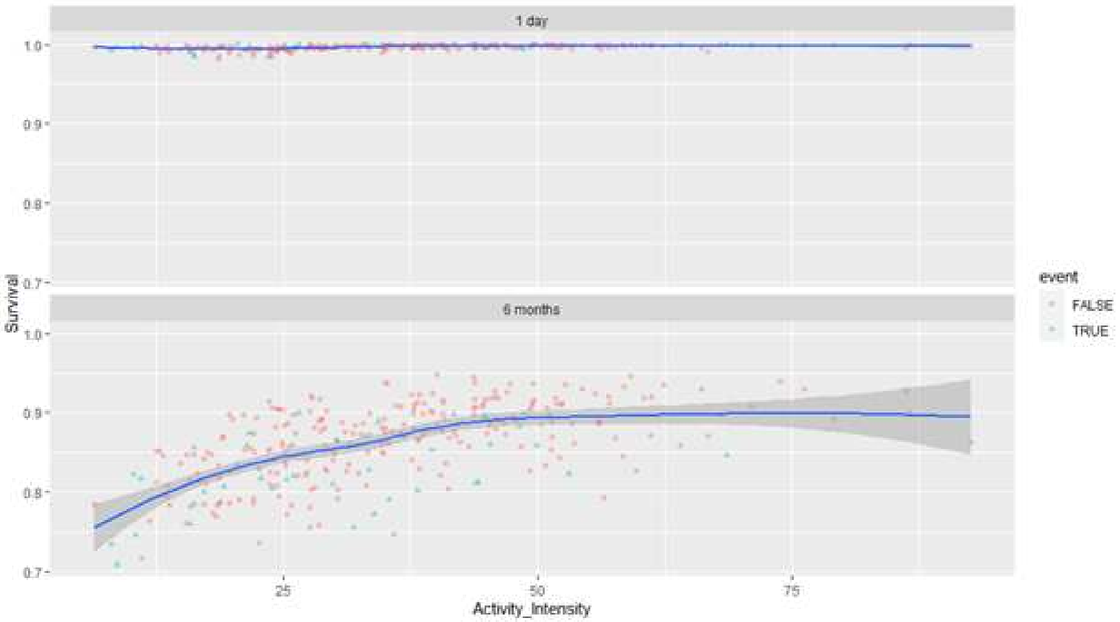


Figure 6.9: Partial dependence plot for activity intensity for clinical outcome of heart failure hospitalization

The function of survival by activity intensity with respect to the clinical outcome of heart failure hospitalization suggests that patients with lower rates of activity intensity, particularly less than an average of 50 m*g*s of sustained activity intensity per day, have lower rates of survival. The function seems to plateau at activity intensity rates greater than 50 m*g*s, however, it is possible this may be due to less patients at the higher levels of activity intensity given that the confidence intervals widen.

6.2 Results for the composite of time to first heart failure hospitalization or urgent heart failure visit based on Random Survival Forest

A random survival forest algorithm was fitted to the independent variables of NT-proBNP at baseline, NYHA class, activity duration, activity intensity, prior heart failure hospitalization, atrial fibrillation, diabetes, age group, sex, EGFR at baseline, BMI at baseline, country group, and hypertension at baseline to predict the composite clinical outcome of time to first heart failure hospitalization and urgent heart failure visit, yielding the below output for the VIMP:

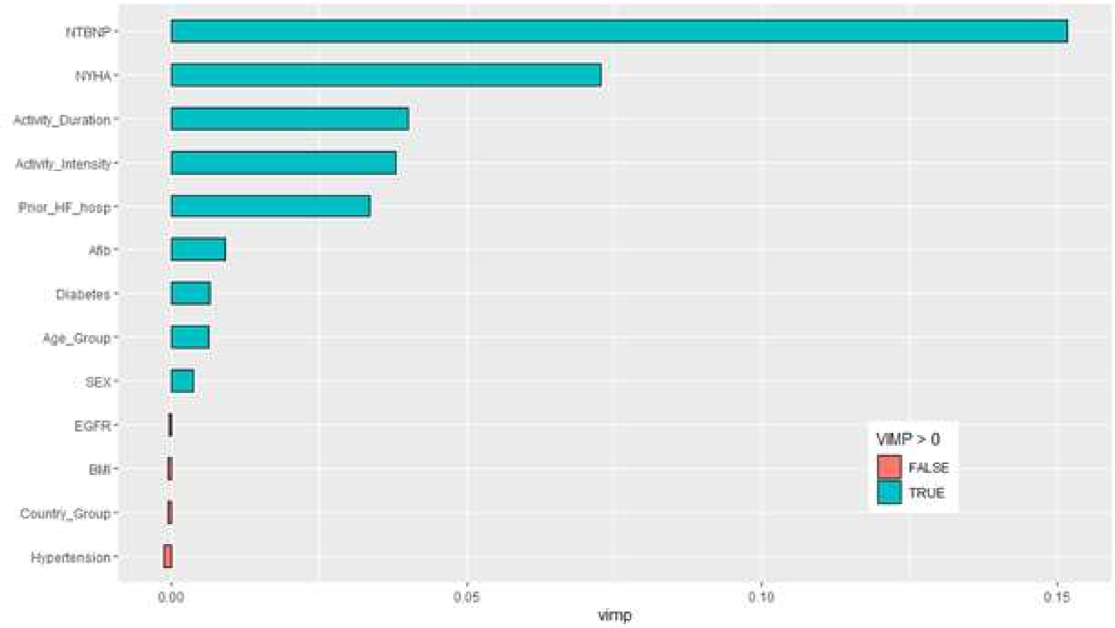
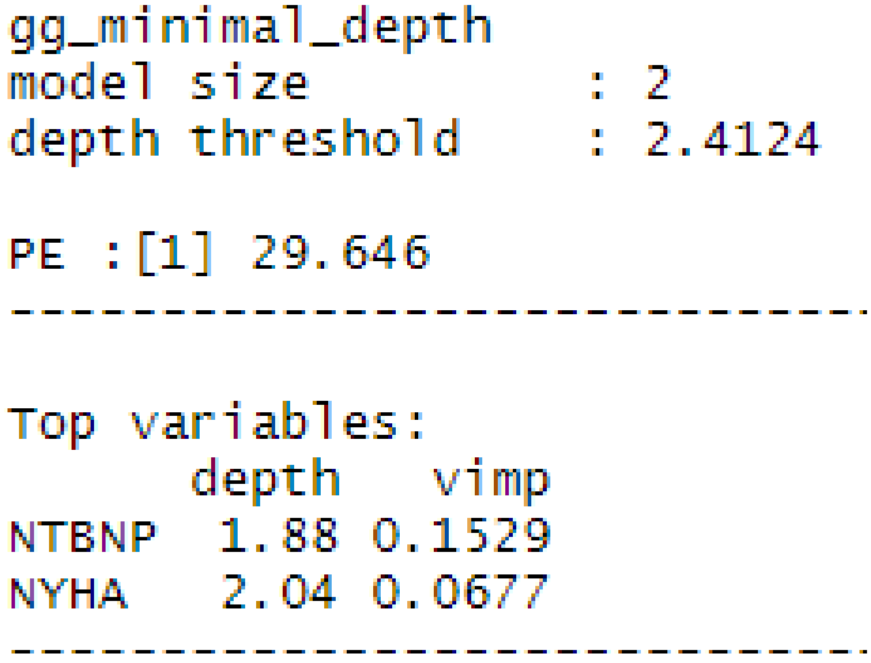


Figure 6.10: Random survival forest VIMP output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit

After following the simulation process grid search denoted above, the model was tuned to the hyperparameters values of nsplit = 2, node size = 20, and mtry = 4, yielding an OOB prediction error for the overall model of 0.2965. The results categorized the variables into two groupings; those with a positive VIMP and those with a negative VIMP. The variables with positive VIMPs for the composite outcome of time to first heart failure hospitalization and urgent heart failure visit included NT-proBNP, NYHA class, activity duration, activity intensity, prior history of heart failure hospitalization, atrial fibrillation, diabetes, age group, and sex. Variables with a negative VIMP included EGFR, BMI, country grouping, and hypertension. Additionally, a minimal depth approach is used to assess the random survival variable importance through another measure to assess similarity of findings. The output for this approach is seen below in graphical and tabular format:

Table 6.2: Random survival forest minimal depth and VIMP scores for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit



```
gg_minimal_depth
model size          : 2
depth threshold     : 2.4124

PE :[1] 29.646
-----------------------------


Top variables:
      depth   vimp
NTBNP  1.88 0.1529
NYHA   2.04 0.0677
-----------------------------
```

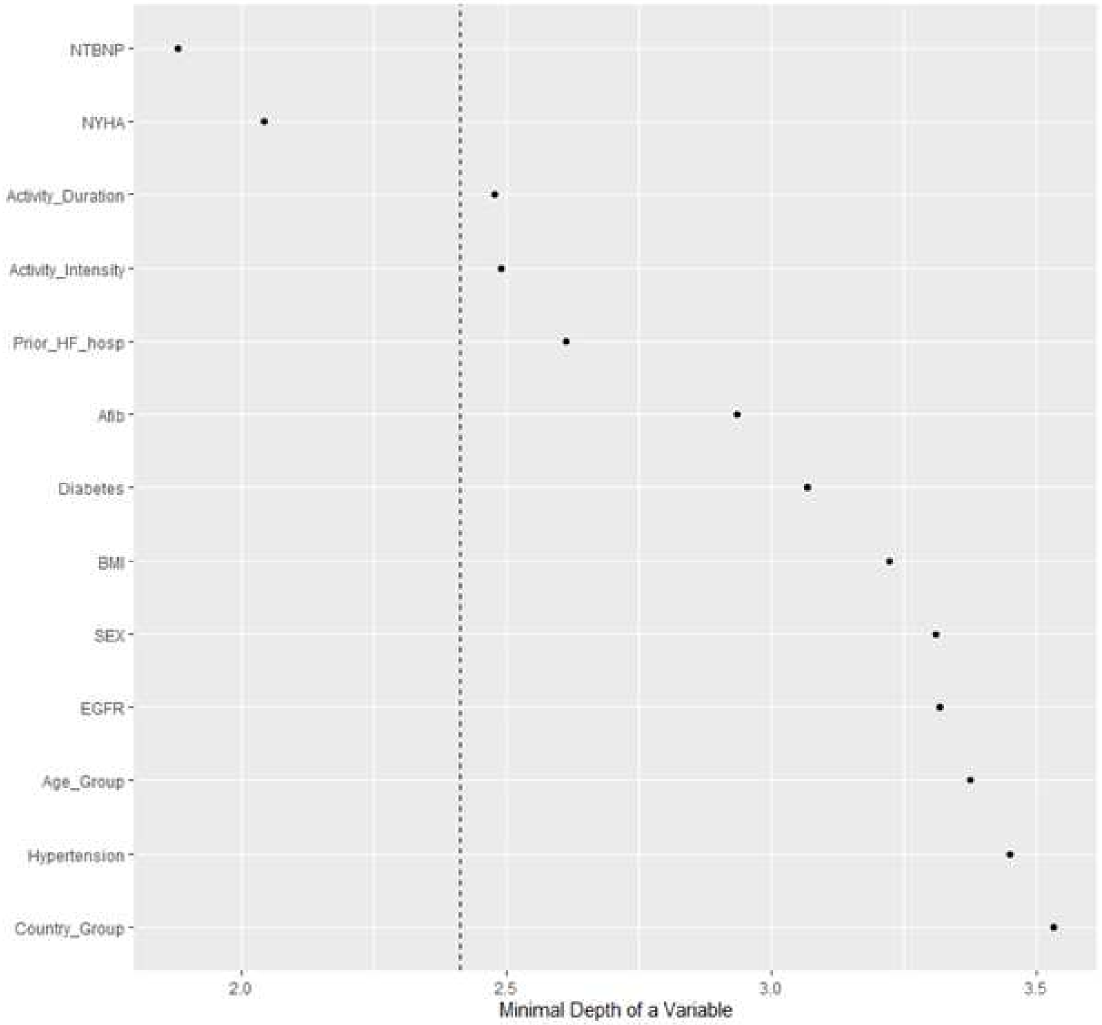


Figure 6.11: Random survival forest minimal depth output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit

The output above indicates that the same top variables were identified by both the VIMP and the minimal depth, including NT-proBNP and NYHA class. However, according to the minimal depth approach activity duration and activity intensity, while the third and fourth most important features as identified by the VIMP, are not identified as important overall as deemed by the minimal depth approach. Plotting both the minimal depth and VIMP together allows for a comparison of the selected variables and their ranking between the two variable importance measures:

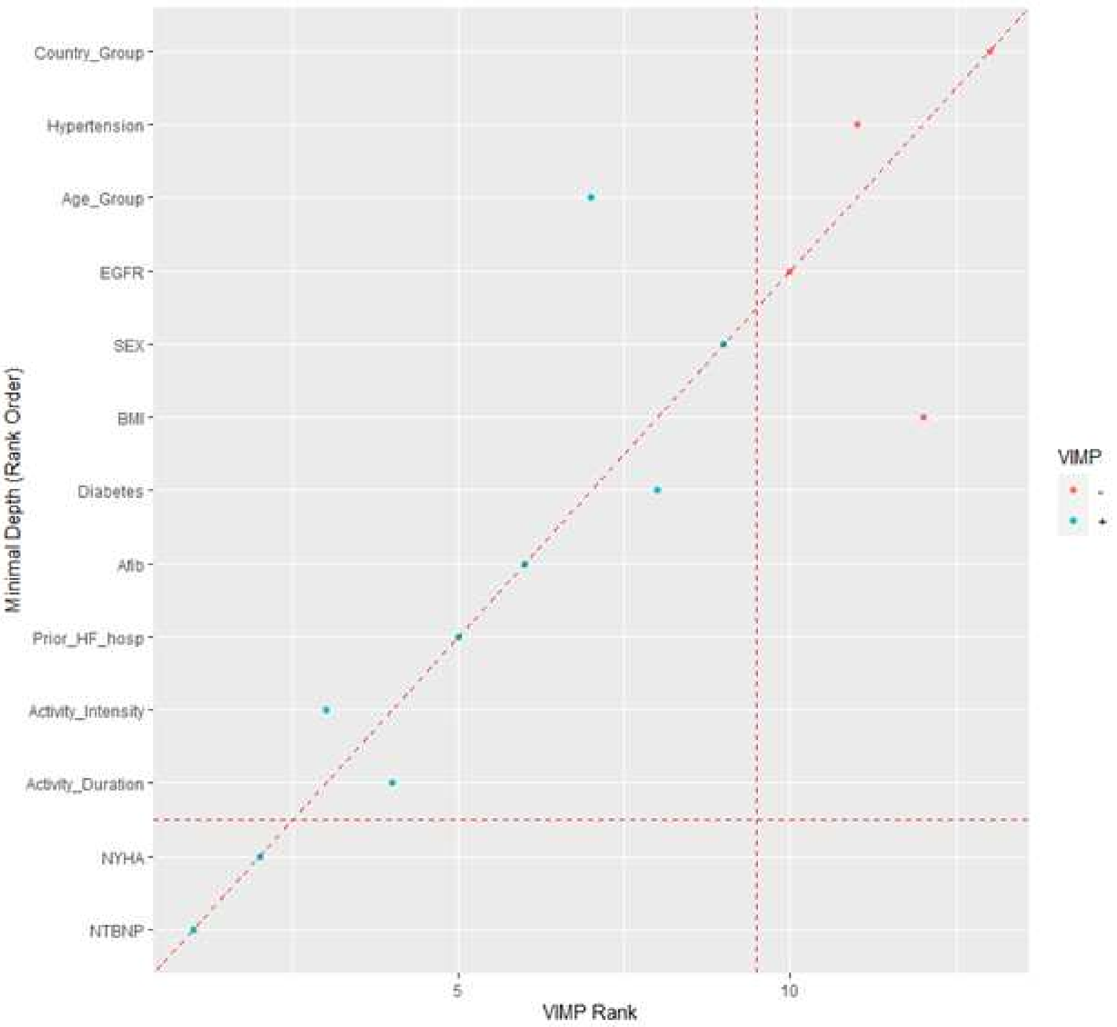


Figure 6.12: Random survival forest minimal depth and VIMP output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit

Concordance between the VIMP and minimal depth is indicated by variables falling on the diagonal on the diagonal line in the plot above, which occurs for most of the variables listed suggesting that the ranking of variables by both variable importance measures is similar. Perfect agreement between the two indices would occur if the variables all fall on the diagonal line. In this case, there is strong overlap between the variables and their importance as measured by the two indices. Given that these variables with VIMP ≤ 0 are not contributing to the predictive power of the model, these variables are removed, and the reduced model is then re-run to assess the appropriateness of including them in the final model. Improvement as quantified by a reduction

OOB error in the reduced model is confirmation that the removal of these variables improves predictive accuracy. The VIMP from the final model with the variables baseline EGFR, BMI, country group, and hypertension removed confirms that all variables left in the reduced model result in the following VIMP plot, now with the variable sex resulting in a negative VIMP:

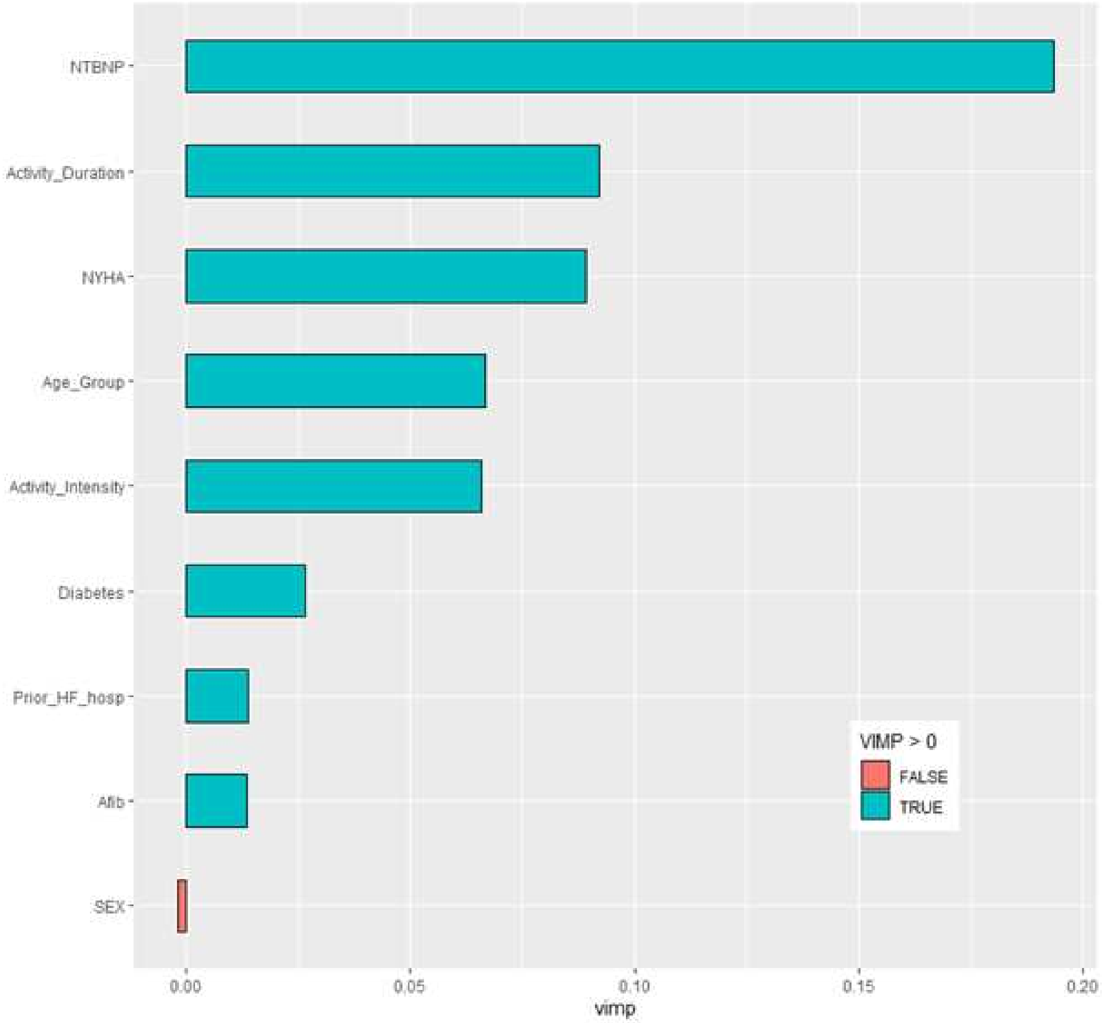


Figure 6.13: Random survival forest intermediate reduced VIMP output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit

The overall model here had a OOB prediction error of 0.284. In an attempt to improve the OOB prediction error, the variable sex was dropped given its negative VIMP and the grid search optimization was re-run yielding the following final reduced model with all remaining predictor variables yielding positive VIMP values:

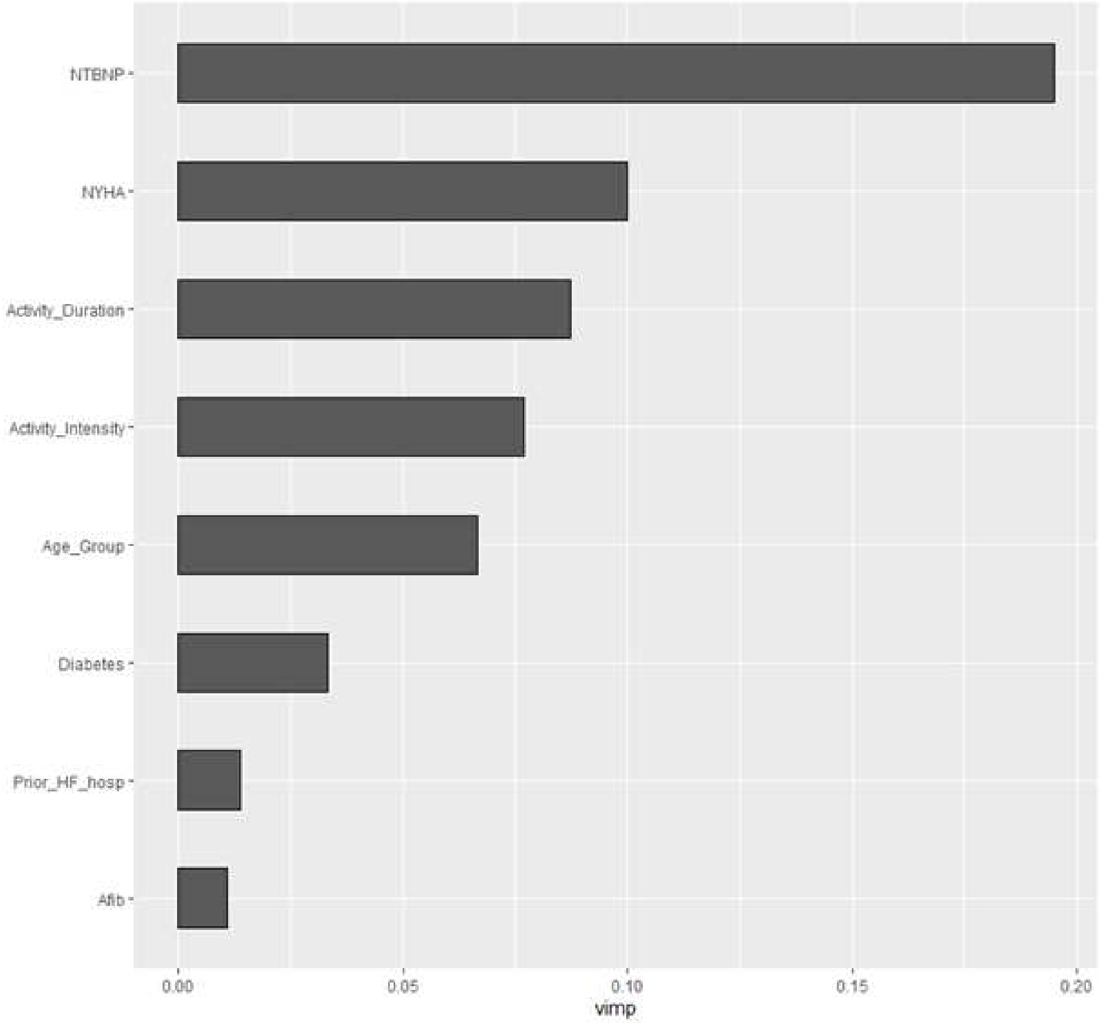


Figure 6.14: Random survival forest final reduced VIMP output for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit

As all of the variables in the above plot have a VIMP > 0, it indicates that all variables in the reduced model contribute to an improvement in the predictive power of the model. This is also confirmed by an improvement in the OOB predictive error which is 0.282 for the final reduced model. Of importance to note here is the ranking of the variables, with NT-proBNP identified as the most important variable in predicting the clinical composite outcome of time to first event of heart failure hospitalization or urgent heart failure visit, followed by NYHA class at baseline, activity duration and then activity intensity. Further examination into the partial dependence plots is performed to better understand the relationship between the variables contributing the most improvement in the predictive power of the model for the clinical composite outcome variable of heart failure hospitalization and urgent heart failure visit.

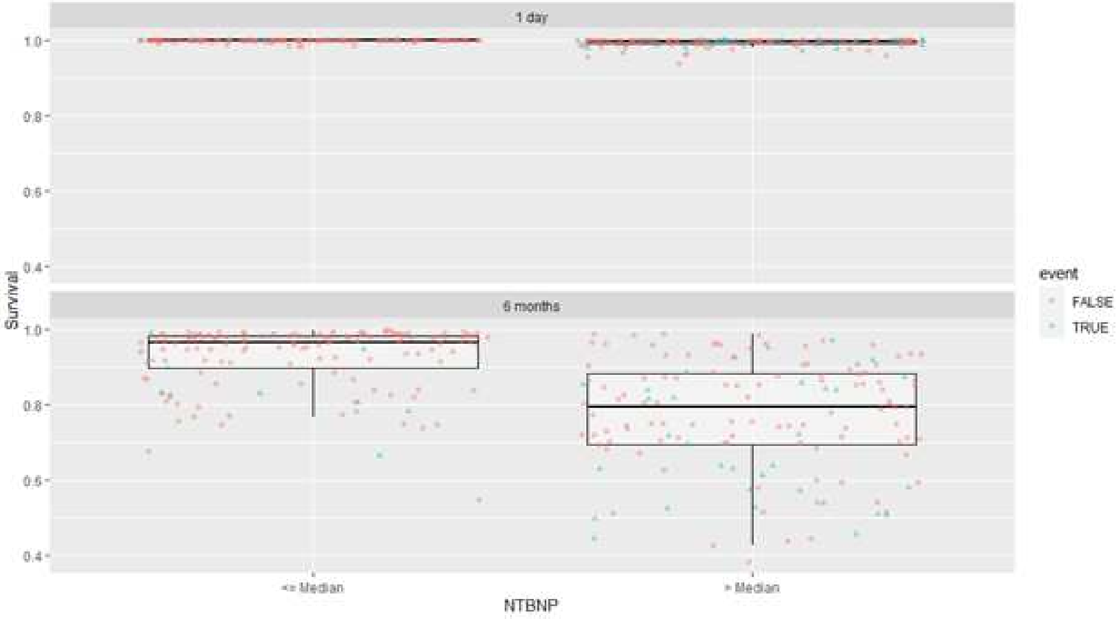


Figure 6.15: Partial dependence plot for baseline NT-proBNP for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit

Patients with baseline NT-proBNP less than or equal to the median have higher survival than patients who have baseline NT-proBNP greater than the median as a predictive variable for the clinical composite outcome of time to first heart failure hospitalization and urgent heart failure visit. Furthermore, fewer patients in the less than or equal to the median baseline NT-proBNP group had the outcome event (11 out of 173 patients or 6.4%) as compared to those in the > median baseline NT-proBNP group (44 out of 174 patients or 25.3%). Baseline NYHA class was identified as the second most important variable and the following partial dependence plot examines survival across the variable:

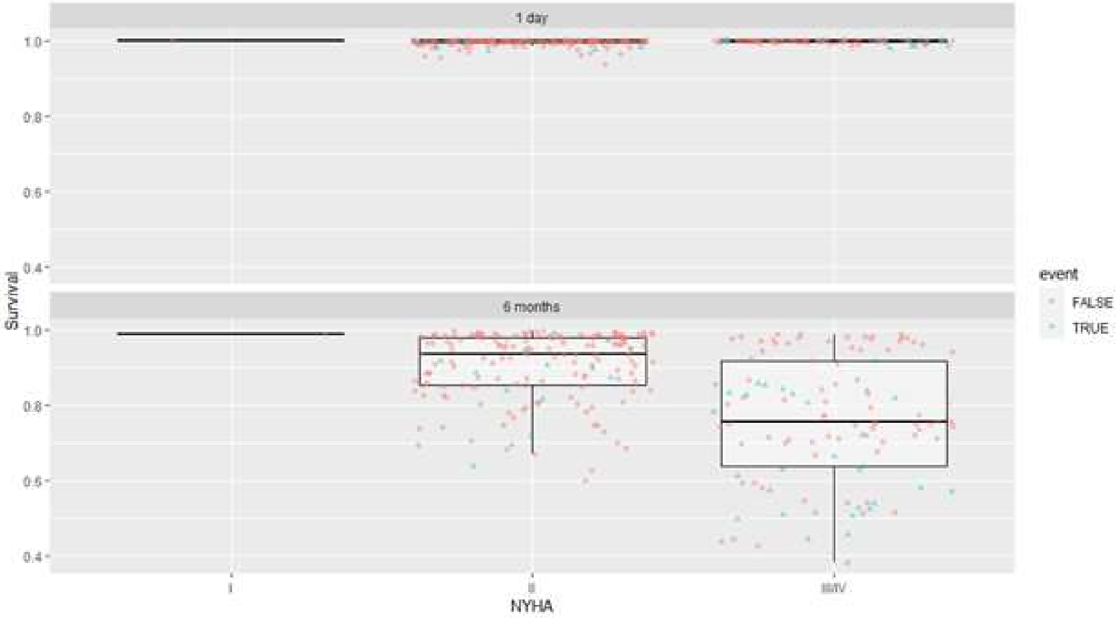


Figure 6.16: Partial dependence plot for baseline NYHA class for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit

Patients in NYHA class I and II had less events and a higher overall survival (19 clinical outcome events out of a total of 208 patients or 9.1%). Patients in NYHA class III and IV had more events and a lower overall survival (36 clinical outcome events out of a total of 139 patients or 25.9%). The relationship of activity duration to survival is examined in the following partial dependence plot:

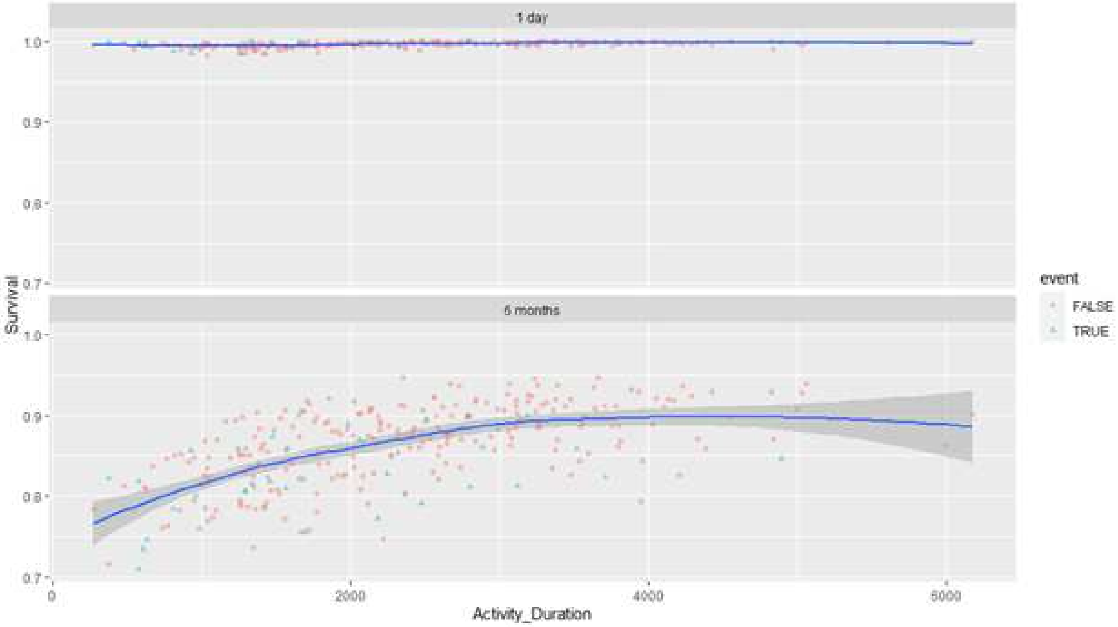


Figure 6.17: Partial dependence plot for baseline activity duration for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit

The function of survival by activity duration with respect to the clinical composite outcome of time to first heart failure hospitalization and urgent heart failure visit suggests that patients with lower rates of activity duration, particularly less than 4000 seconds or approximately 67 minutes per day, have lower rates of survival. The function seems to plateau at activity duration rates greater than 4000 seconds or approximately 67 minutes per day, however, it is possible this may be due to less patients at the higher levels of activity duration given the increasing width of the confidence intervals at that point in the plot. The relationship of activity intensity to survival is examined in the following partial dependence plot:

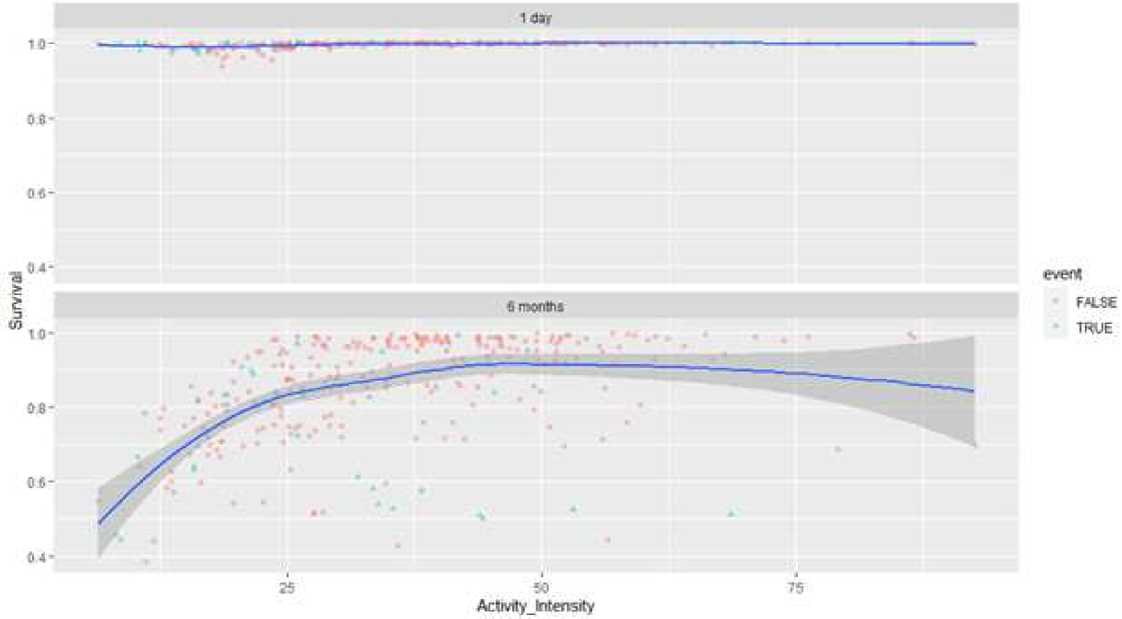


Figure 6.18: Partial dependence plot for baseline activity intensity for clinical outcome of the composite of time to first heart failure hospitalization or urgent heart failure visit

The function of survival by activity intensity with respect to the clinical composite outcome of time to first heart failure hospitalization and urgent heart failure visit suggests that patients with lower rates of average activity intensity, particularly less than an average of 40 m*g*s of sustained activity intensity per day, have lower rates of survival. The function seems to plateau at activity intensity rates greater than 40 m*g*s, however, it is possible this may be due to less patients at the higher levels of activity intensity given the width of the confidence intervals at this point in the plot.

6.3 Results for the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death based on Random Survival Forest

A random survival forest algorithm was fitted to the independent variables of NT-proBNP at baseline, NYHA class, activity duration, activity intensity, prior heart failure hospitalization, atrial fibrillation, diabetes, age group, sex, EGFR at baseline, BMI at baseline, country group, and hypertension at baseline to predict the dependent clinical outcome variable of time to first event of for the composite of time to first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death, yielding the below output for the VIMP:

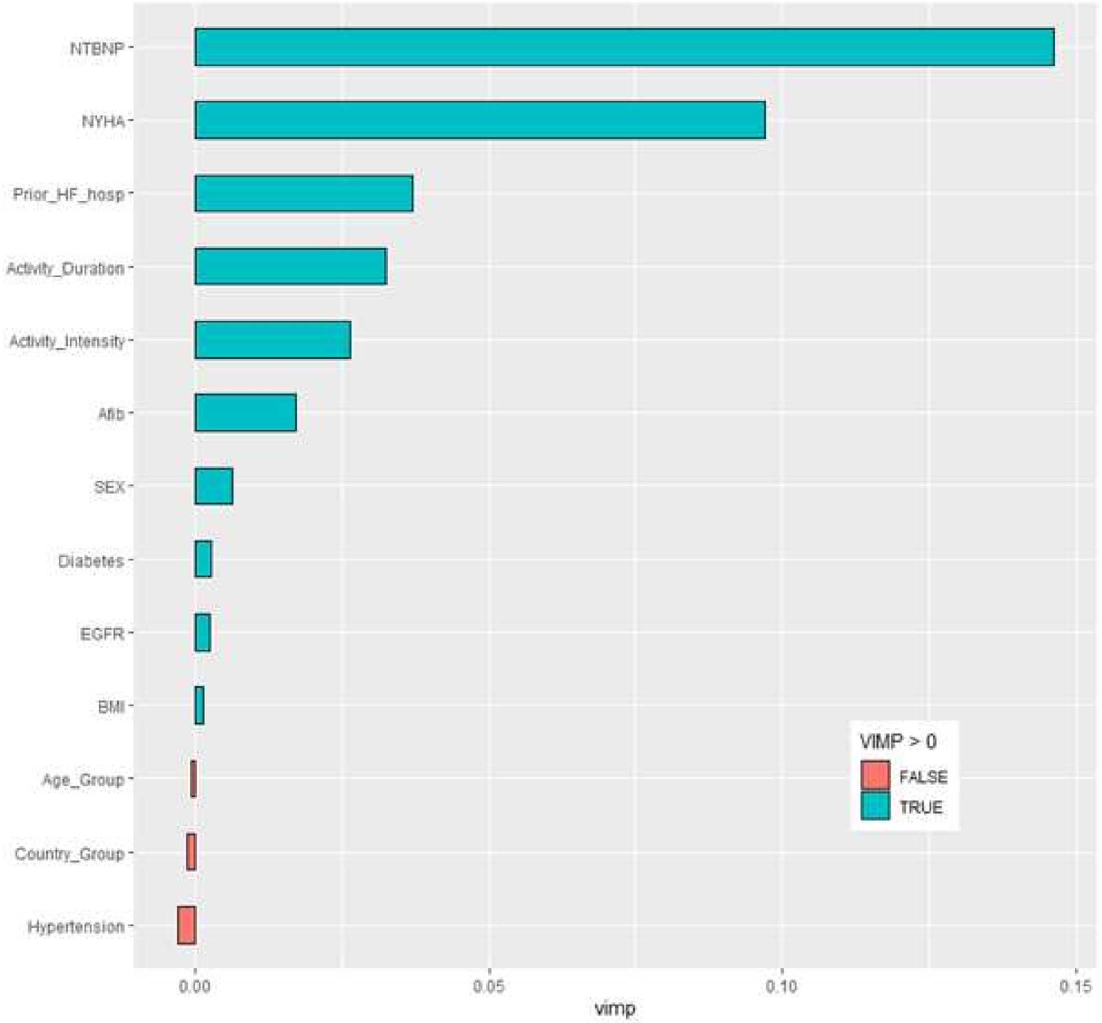
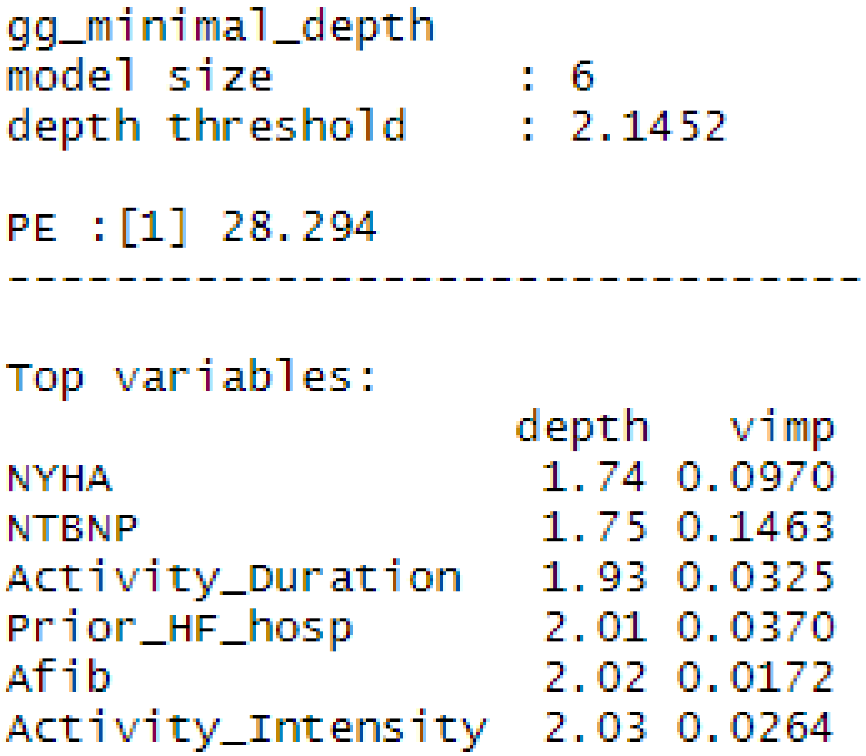


Figure 6.19: Random survival forest VIMP output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death

After following the simulation process grid search denoted above, the model was tuned to the hyperparameters values of nsplit = 2, node size = 35, and mtry = 3, yielding an OOB prediction error for the overall model of 0.2829. The results categorized the variables into two groupings;

those with a positive VIMP and those with a negative VIMP. The variables with positive VIMPs for the outcome of time to first event of the composite of time to first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death included NT-proBNP, NYHA class, prior history of heart failure hospitalization, activity duration, activity intensity, atrial fibrillation, sex, diabetes, EGFR, and BMI. Variables with a negative VIMP included age grouping, country grouping, and hypertension. Additionally, a minimal depth approach is used to assess the random survival variable importance through another measure to assess similarity of findings. The output for this approach is seen below in graphical and tabular format:

Table 6.3: Random survival forest minimal depth and VIMP scores for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death



```
gg_minimal_depth
model size          : 6
depth threshold     : 2.1452

PE :[1] 28.294
----------------------------------

Top variables:
                    depth   vimp
NYHA                 1.74 0.0970
NTBNP                1.75 0.1463
Activity_Duration    1.93 0.0325
Prior_HF_hosp        2.01 0.0370
Afib                 2.02 0.0172
Activity_Intensity   2.03 0.0264
```

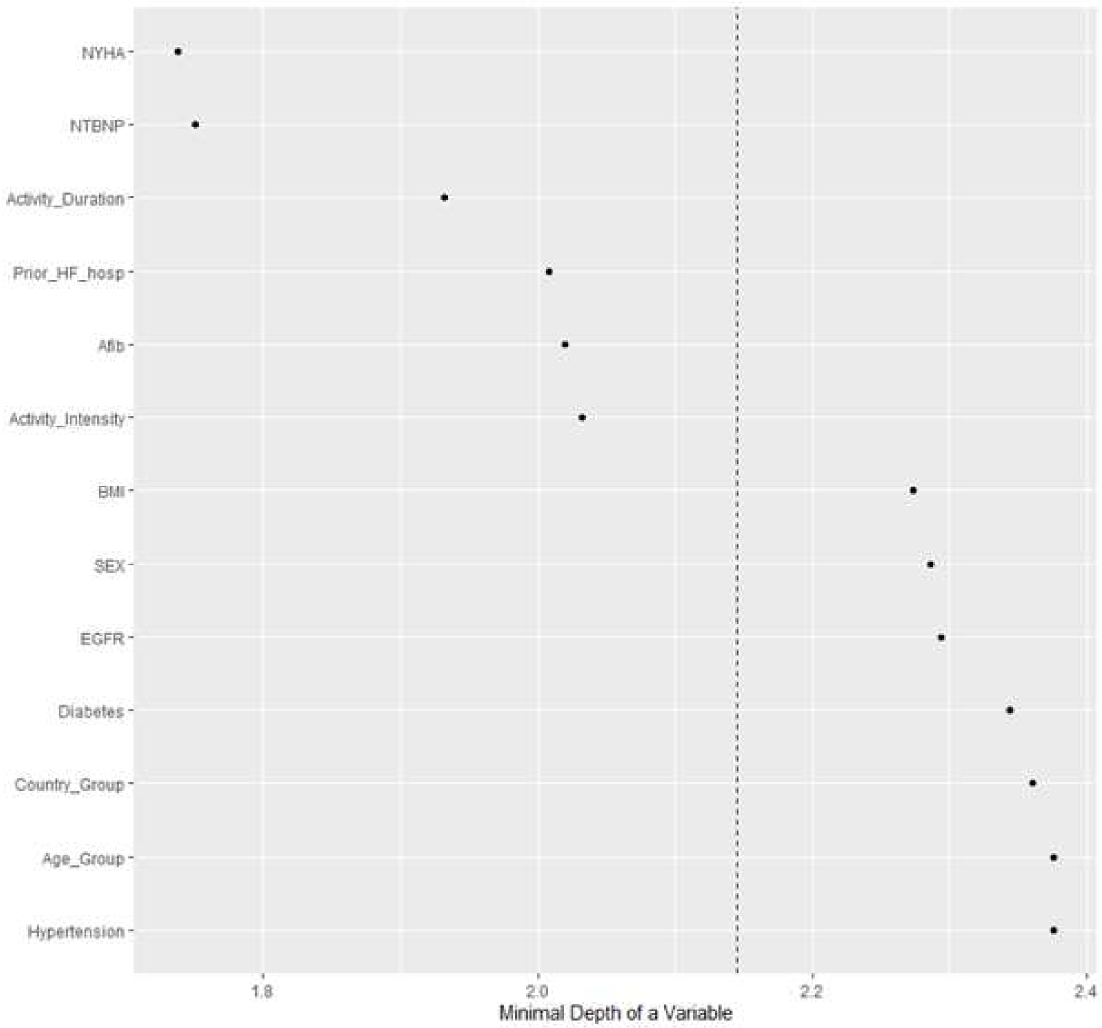


Figure 6.20: Random survival forest minimal depth output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death

The output above indicates that the same top variables were identified by both the VIMP and the minimal depth, including NYHA class, NT-proBNP, activity duration, prior heart failure hospitalization, atrial fibrillation, and activity intensity. However, according to the minimal depth approach, BMI, sex, EGFR, and diabetes are not identified as important overall as deemed by the algorithm. Plotting both the minimal depth and VIMP together allows for a comparison of the selected variables and their ranking between the two variable importance measures:

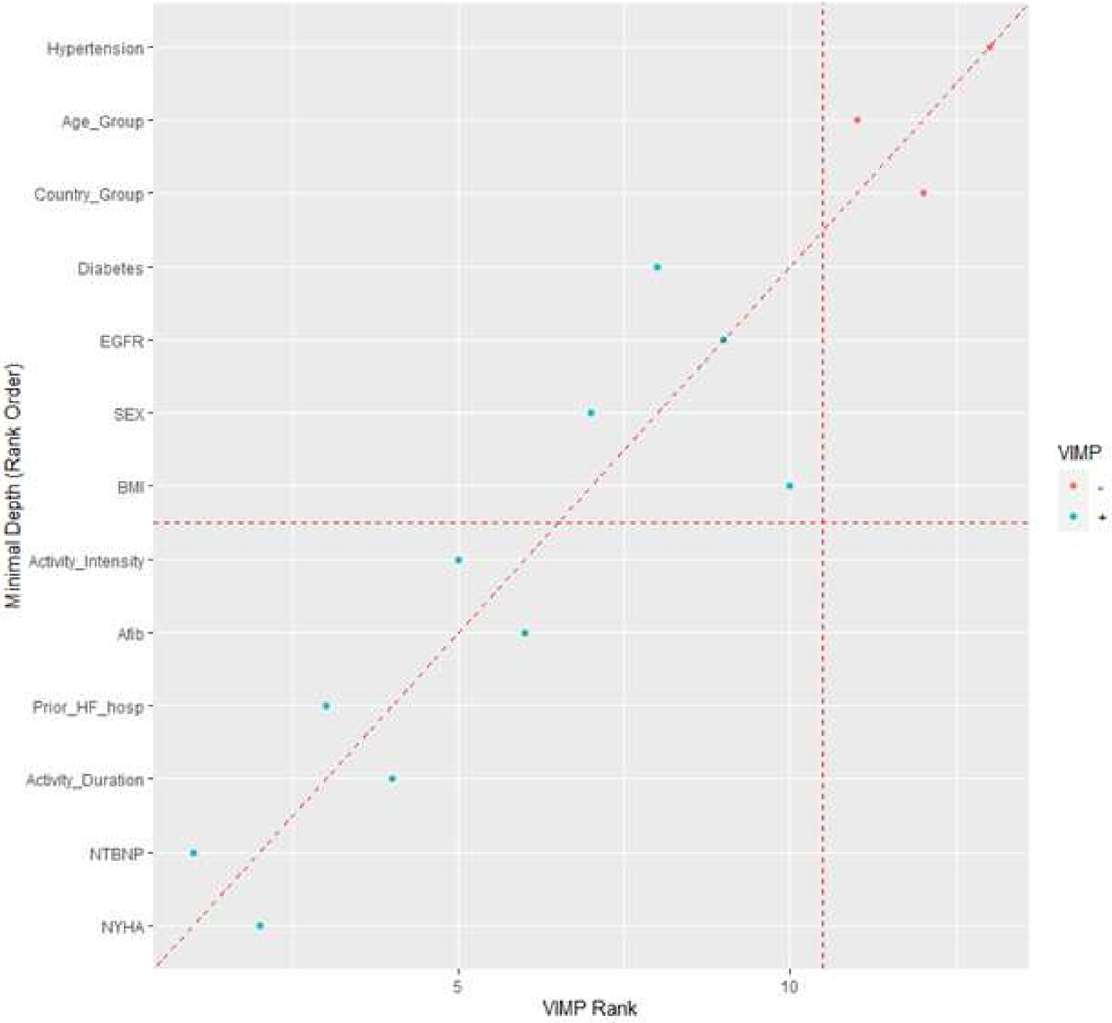


Figure 6.21: Random survival forest minimal depth and VIMP output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death

Concordance between the VIMP and minimal depth is indicated by variables falling on the diagonal on the diagonal line in the plot above, which occurs for most of the variables listed suggesting that the ranking of variables by both variable importance measures is similar. Perfect agreement between the two indices would occur if the variables all fall on the diagonal line. In this case, there is strong overlap between the variables and their importance as measured by the two indices. There are several variables, including age group, country group, and hypertension that yielded VIMP scores ≤ 0, suggesting that these variables are not contributing to the predictive power of the model. These variables are removed and the reduced model is re-run to assess the appropriateness of excluding them in the final model. Improvement as quantified by a reduction OOB error in the reduced model is confirmation that the removal of these variables improves predictive accuracy. The VIMP from the reduced model with the variables age group, country group, and hypertension removed confirms that most of the remaining variables left in the reduced model result in a positive VIMP, however, now the variables BMI and baseline EGFR are yielding a VIMP score ≤ 0.

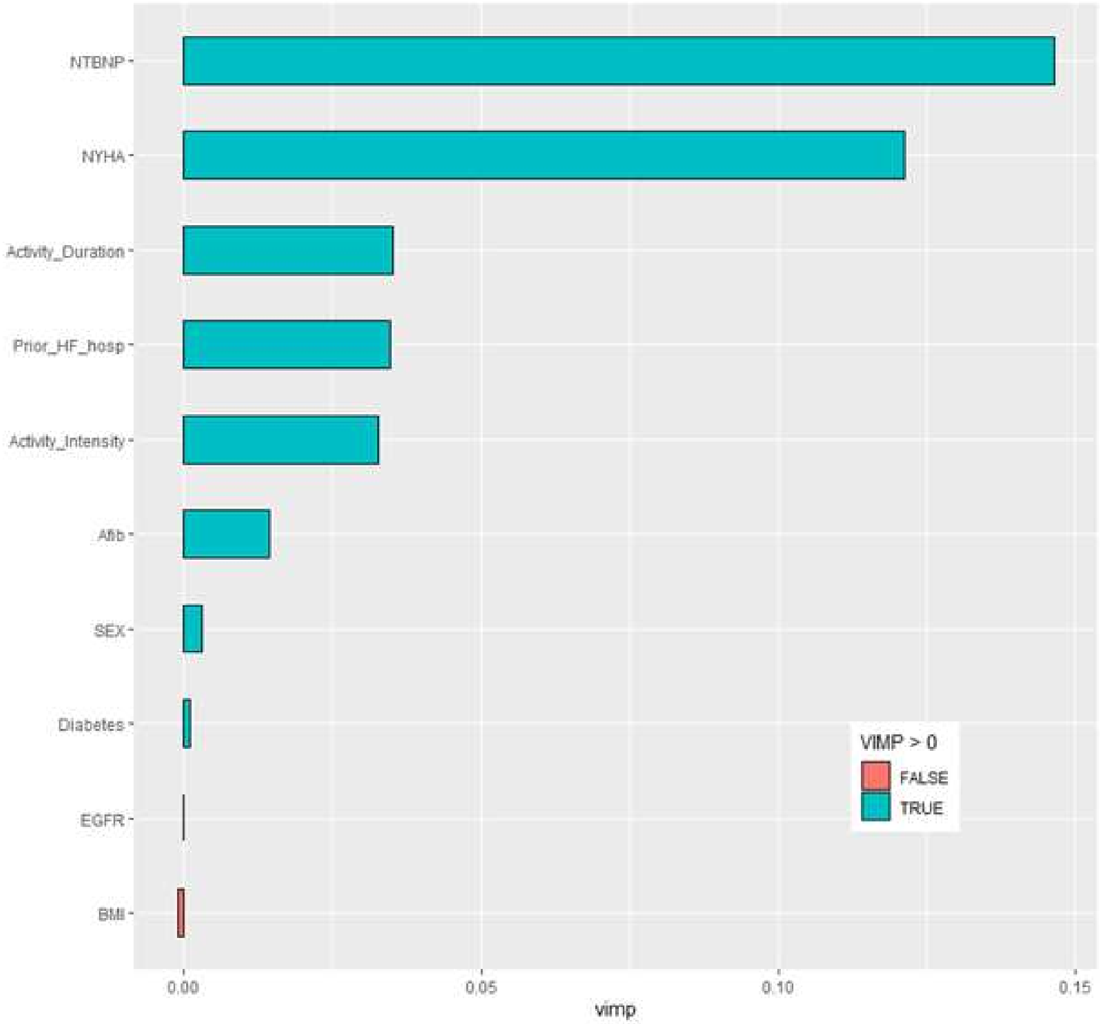


Figure 6.22: Random survival forest intermediate reduced VIMP output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death

This intermediate reduced model plotted above yields an OOB prediction error of 0.278, suggesting an improvement from the original model containing all variables. In an attempt to improve the model further, the grid search simulation was re-run with the variables yielding a VIMP ≤ 0 removed (baseline EGFR and BMI), and the following final reduced model was identified with all remaining predictor variables resulting in positive VIMP values:

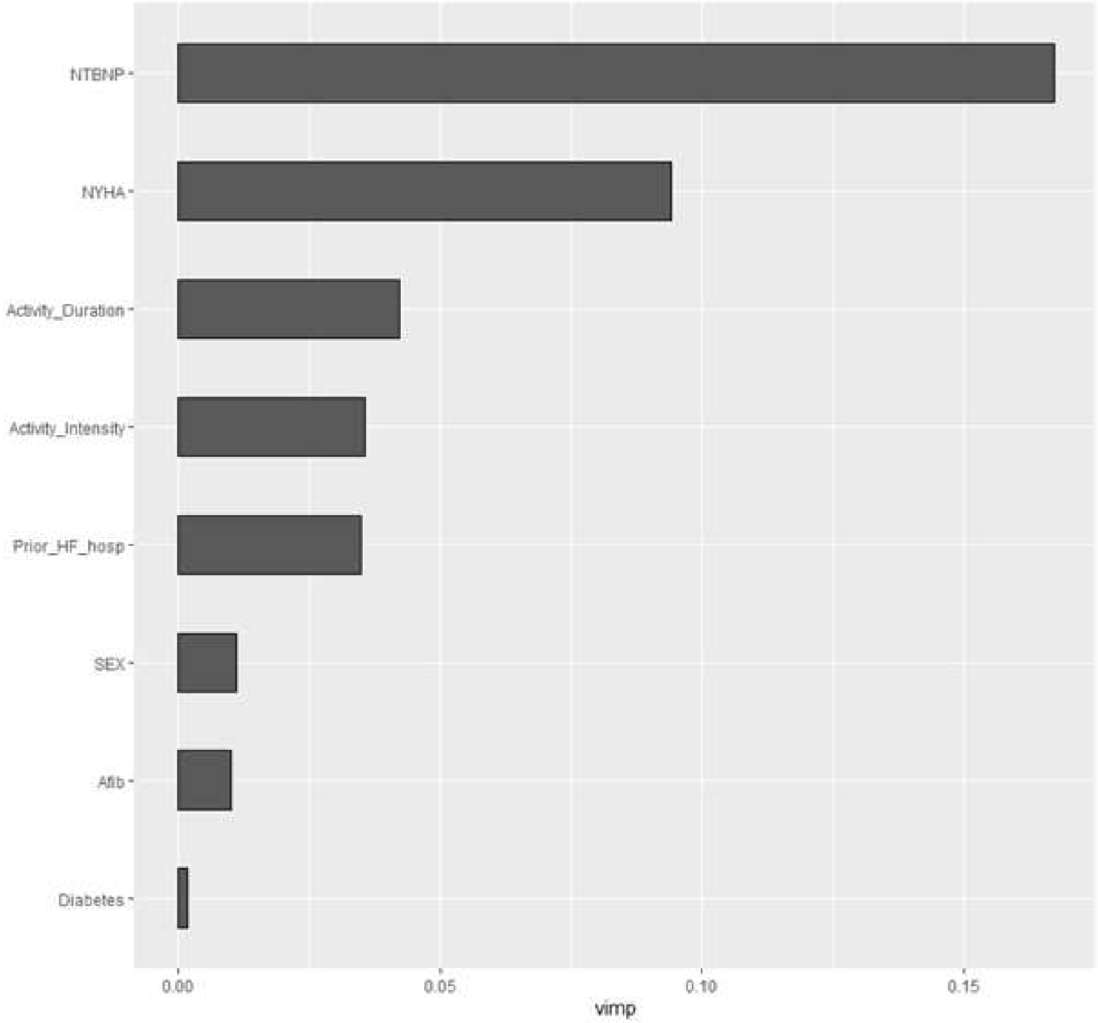


Figure 6.23: Random survival forest final reduced VIMP output for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death

As all of the variables in the final reduced model have a VIMP > 0, it indicates that all variables in the model contribute to an improvement in the predictive power of the model. This is also confirmed quantitatively by an improvement in the OOB predictive error which is 0.2732 for the model. Of importance to note here is the ranking of the variables, with NT-proBNP identified as the most important variable in predicting the clinical outcome of the composite of time to first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death, followed by NYHA class at baseline, activity duration and then activity intensity. Further examination into the partial dependence plots is performed to better understand the relationship between the variables contributing the most improvement in the predictive power of the model and the clinical outcome variable of the composite of time to first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death.

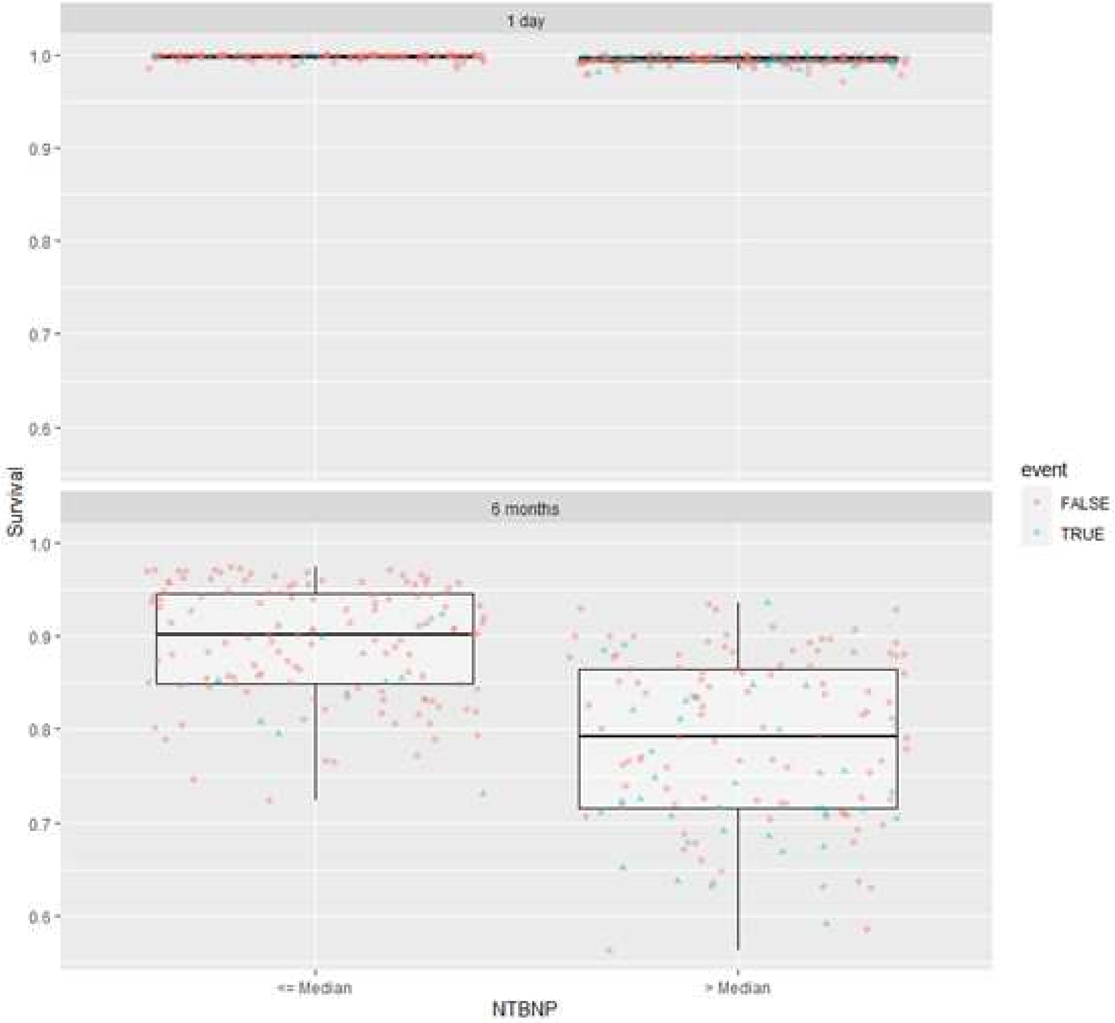


Figure 6.24: Partial dependence plot for baseline NT-proBNP for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death

Patients with baseline NT-proBNP less than or equal to the median have higher survival than patients who have baseline NT-proBNP greater than the median as a predictive variable for the clinical outcome of the composite of time to first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death. Furthermore, fewer patients in the less than or equal to the median baseline NT-proBNP group had the outcome event (11 out of 173 patients or 6.4%) as compared to those in the greater than the median baseline NT-proBNP group (47 out of 174 patients or 27.0%). Baseline NYHA class was identified as the second most important variable and the following figures examines survival across the variable:

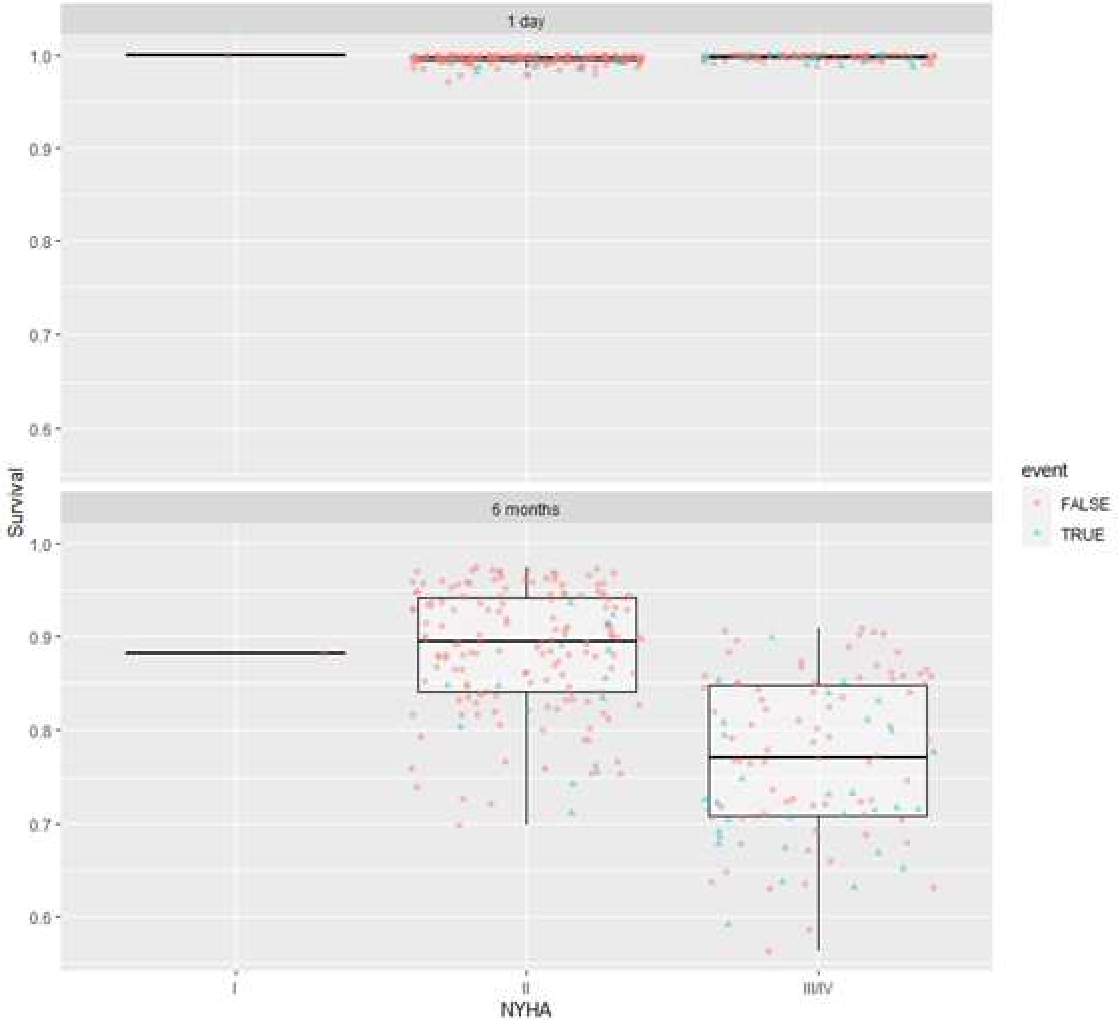


Figure 6.25: Partial dependence plot for baseline NYHA class for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death

Patients in NYHA class I and II had less events and a higher overall survival (19 clinical outcome events out of a total of 208 patients or 9.1%). Patients in NYHA class III and IV had more events and a lower overall survival (39 clinical outcome events out of a total of 139 patients or 28.1%). The relationship of activity duration to survival is examined in the following partial dependence plot:

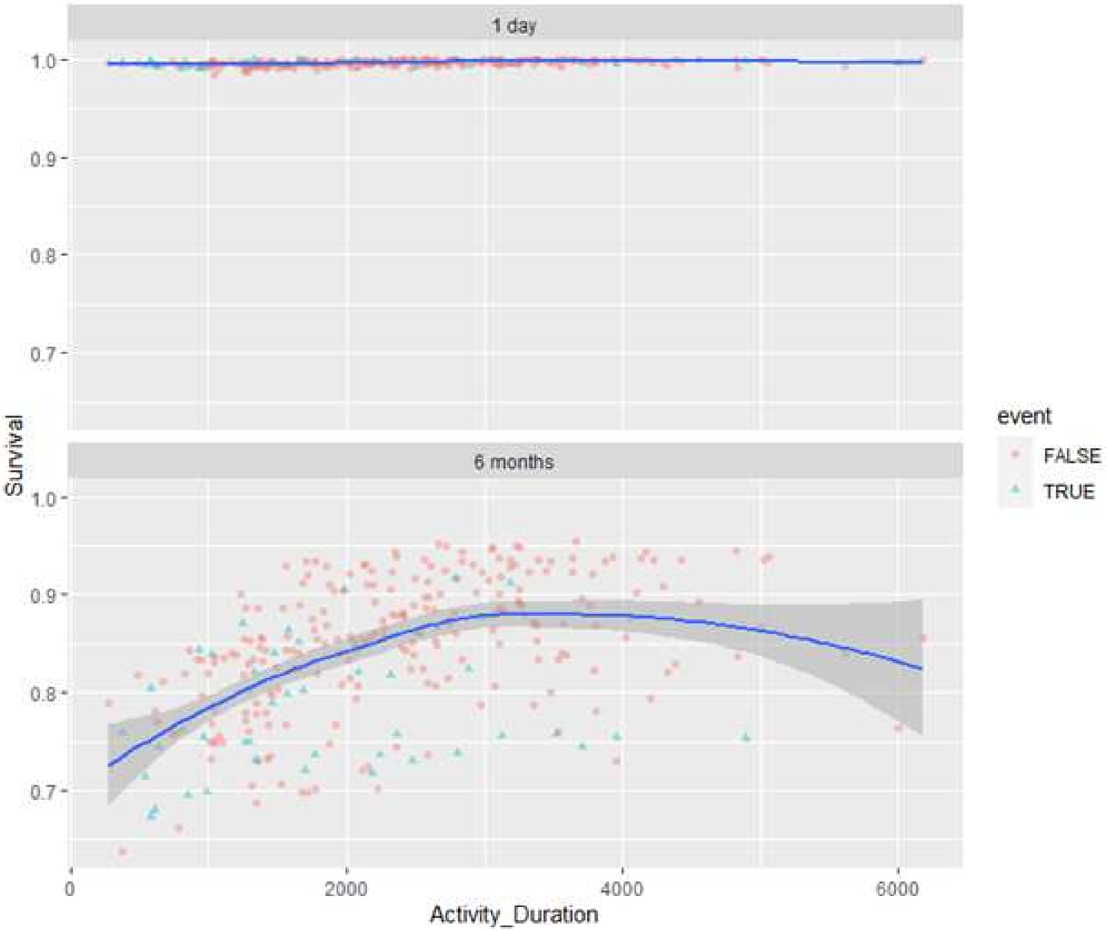


Figure 6.26: Partial dependence plot for baseline activity duration for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death

The function of survival by activity duration with respect to the clinical outcome of the composite of time to first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death suggests that patients with lower rates of activity duration, particularly less than 3000 seconds or 50 minutes per day, have lower rates of survival. The function seems to plateau at activity duration rates greater than 3000 seconds or 50 minutes per day, however, it is possible this may be due to less patients at the higher levels of activity duration given that the confidence intervals widen at this point in the plot. The relationship of activity intensity to survival is examined in the following partial dependence plot:

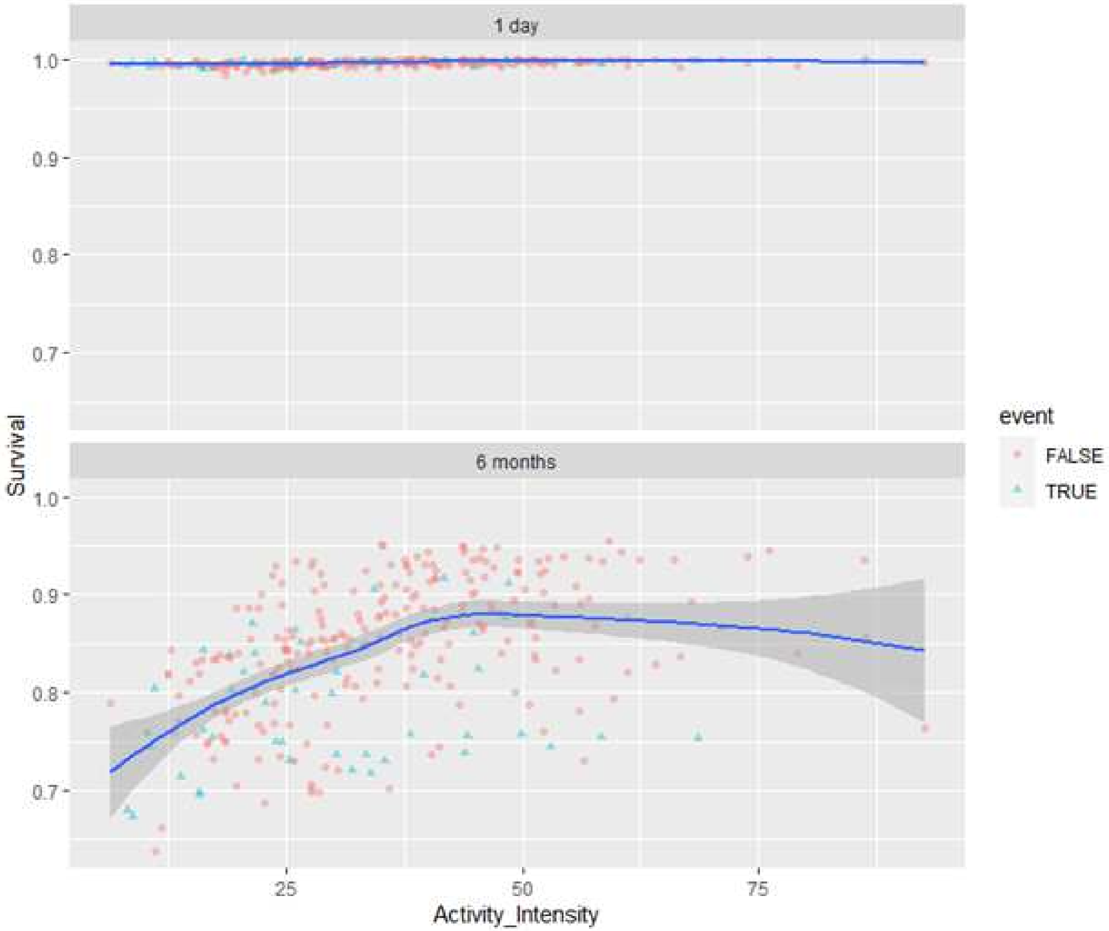


Figure 6.27: Partial dependence plot for baseline activity intensity for clinical outcome of the composite of time to first heart failure hospitalization, urgent heart failure visit, or cardiovascular death

The function of survival by activity intensity with respect to the clinical outcome of the for the composite of time to first event of heart failure hospitalization, urgent heart failure visit, or cardiovascular death suggests that patients with lower rates of average activity intensity, particularly less than 40 m*g*s of sustained activity per day, have lower rates of survival. The function seems to plateau at average activity intensity rates greater than 40 m*g*s, however, it is possible this may be due to less patients at the higher levels of activity intensity given that the confidence intervals widen at this point of the plot.

6.4 Discussion for Random Survival Forest

Overall, the RSF models yielded similar strong predictive accuracy with respect to OOB prediction error (0.2987, 0.2820, 0.2732 for the three reduced models corresponding to the three clinical outcomes, respectively). As mentioned previously, an OOB predictive error of 50% indicates that the model is performing at chance. Given the small dataset and the variability in biological data, the OOB prediction error values generated by the models presented here can be considered as reflecting useful models with strong performance. Based on the model characteristics and performance, it may be possible to use these models to identify patients who are in earlier stages of their disease progression and introduce therapeutics that will reduce the likelihood of progression to later stages.

Additional, one may be able to use this improved understanding of physical activity data and its association to clinical outcomes to identify patients earlier on in their disease progression and improve on this area of unmet medical need. This identification can subsequently be used to provide patients with an earlier treatment intervention. Based on recent surveys of US adults in 2021, Pew Research Center has identified that 85% of adults in America have smartphones [54], who would thus have the capability of having their physical activity and intensity data assessed. Assuming the earlier mentioned approximately 5.7 million patients suffering from heart failure [9], a potential 85% of them or approximately 4.85 million would have access to a smartphone. Coupled with the utility of the proposed activity models, these patients could then utilize their smartphone to track their physical activity data and have it assessed to predict the likelihood of the individual patient experiencing an event. This would have a significant improvement in cost and outcomes for the healthcare system, improving on things like general efficiency, reducing cost, and improving patient's quality of life by identifying heart failure decompensations earlier before they result in outcomes like myocardial infarction or cardiovascular death.

Furthermore, it is interesting to note that all 3 RSF models further implicated NT-proBNP and NYHA class as the most important predictor variables, similar to what was identified by the Cox PH models in the previous section. According to the partial dependence plots, on average and across the models, patients in the NT-proBNP group with baseline values $\leq$ to the median had a 90% survival at 6 months as compared to those in the > median NT-proBNP group who had only a 77% median survival at 6 months. This is in line with previously discussed information that NT-proBNP plays an important role as a diagnostic biomarker variable, with higher values of NT-proBNP being associated with more recent decompensations in heart failure [7]. Correspondingly, these higher values are associated with higher rates of the clinical outcome events, as can be seen in the differential survival of 12% amongst these two groups. On average and across the models, NYHA class patients in class II had 89% median survival at 6 months, compared to 77% median survival in class III/IV. This is further supported by the literature stating that patients in class III/IV represent more severe heart failure patients with an increased risk for clinical outcomes [25]. This finding also corresponds well with the importance of the activity intensity and activity duration variables as identified by the RSF models. Through examination of the partial dependence plots, it is evident that higher rates of activity intensity and activity duration yield improvements in median survival times over the 6-month period. In particular, those patients who are able to engage only in lower levels of activity intensity and activity duration have strikingly lower survival probabilities across all three models. Taken collectively, the information from the RSF models suggests that the variables of NT-proBNP, NYHA class, activity intensity, and activity duration are particularly important and provide positive predictive performance across the three clinical outcomes assessed.

Chapter 7

Discussion and clinical interpretation of findings

The present work provides a comprehensive overview of topics to consider when analyzing wearable and biosensor data, particular data from accelerometers used to assess physical activity in patients. In the case where one is interested in understanding how patients relate to one another without taking into account clinical outcomes data, perhaps because the trial has just begun and only baseline values are available, k-means clustering and consensus clustering algorithms are presented to help categorize the patients into groupings. These clusters can then be used to help examine any heterogeneity across patients clusters and profiles and to identify if certain patient groups are responding differently within the trials. In the case where one is interested in developing models to predict clinical outcomes of patients, the Cox PH model and random survival forest approaches were presented to include approaches able to predict the time-to-event clinical outcome data based on independent variables including clinical data and accelerometry wearable device data. In these cases, the wearable device data proved complimentary to other strong predictive variables such as NT-proBNP and NYHA class and helped to improve model performance. While it is evident that there are many different variables that are implicated in the understanding of patients, their journeys, clinical outcomes, and general patient quality of life, the analytical assessments presented here suggest that certain variables, such as NT-proBNP and NYHA class, can provide strong ability to differentiate between patients and predict their outcomes. Furthermore, it is suggested by the results of the clustering algorithms and predictive models that physical activity data from biosensors and wearable devices provide informative complimentary information for the models to utilize that allows the algorithms to aggregate the patients into clinically relevant and meaningful clusters as well as improve predictive accuracy for clinical outcome models.

In summary, the overall findings across the algorithms utilized suggested that a signal emerged from the activity duration and activity intensity data in that patients who engaged in lower levels of physical activity and activity intensity represented a clinically relevant differentiated group of heart failure patients from those heart failure patients who were able to engage in higher levels of physical activity duration and activity intensity. These patients with lower levels of activity intensity and duration were generally older, had more previous serious comorbidities, had lower EGFR suggesting kidney impairment, higher NT-proBNP values at baseline, and were more likely to be categorized as NYHA class III/IV patients compared to their higher activity engaging counterparts. Not only did these patients suffer from these additional negative associations within their baseline data, but these patients also had a lower probability of survival as indicated in the predictive model assessments.

While physical limitation and a general negative impact on physical ability and quality of life is a common complaint of patients suffering from cardiovascular disease and heart failure, the ability to assess a patient's physical capabilities in everyday life have not been possible given technical limitations in the development of and utilization of wearables and biosensors in clinical trials.

While assessments on quality of life can be performed through the use of surveys and questionnaires that represent subjective assessments of physical limitation, a wearable can be utilized to examine a patient's physical capabilities over the course of the trial and as indicated by the results of this work, can provide meaningful complimentary information to the understanding of the patient's likelihood for clinical outcomes. As mentioned previously, utilizing physical activity data from the smartphones phones that many American adults regularly carry, could potentially result in nearly 4.85 million heart failure patients having their physical activity data modeled. This data could be used to improve the predictive accuracy of patient outcome models based on physical activity and to potentially classify patients and provide treatment interventions earlier on. Given the median cost of heart failure hospitalization is \$ 13,418 [13], even a small improvement in reducing heart failure hospitalizations could potentially dramatically reduce the cost of heart failure treatment. Optimistically, if the physical activity data was routinely assessed and utilized in models from all heart failure patients who used smartphones in the US, the upper limit of potential treatment costs saved solely for heart failure hospitalization would eclipse \$ 65 billion.

Given that the heart failure population examined here represents a very compromised patient population that has strong limitations in physical activity, it is possible that a feature such as activity duration may be more sensitive than a feature such as activity intensity, as patients are often only able to engage in lower levels of activity intensity across the majority of their data. One would need to adapt this when examining healthier patients, as it is possible additional variables would be necessary to further differentiate aspects of physical activity deterioration and improvement form one another.

7.1 Limitations and future development

There are some general limitations to the work presented here. One clear and main limitation is the small sample size utilized. While the results of the work are biologically plausible and did yield support based on evidence reported in the literature, it is still questionable how well the results would generalize to other heart failure patients. In particular, given the small proportion of females enrolled in the trial, one would need to further assess the results in this population before generalizing it to a wider female heart failure population. Furthermore, given that the scope of guideline directed medical therapy for heart failure is ever changing, it is possible that clinically relevant clusters identified in trials from 5-10 years ago may be different today based on things such as medical standard of care variability or refinement in inclusion / exclusion criteria. Nevertheless, the work here can serve as a foundation to expanding upon some of the future work in these areas. A question that still needs to be addressed within the scientific literature is what resolution of wearable and biosensor data is appropriate to assess changes in patients, and more specifically in the area of physical activity. While the focus on these assessments were based on an hourly level of resolution aggregated to a daily level, it is possible that even daily summaries would still capture the signal necessary to differentiate patients. Nevertheless, this question remains open and would need to be answered experimentally to fully understand the repercussions of utilizing different resolutions. One can envision the appropriate resolution to analyze data would be dependent on the patient population being studied. While wearable and biosensor data at a higher resolution can necessitate a large amount of storage space, one can always then summarize the data to a lower resolution using something like a moving average approach, whereas it is impossible to accurately backtrack the data from a lower resolution to a higher resolution.

While the work presented here reflected several workstreams within wearable devices and biosensors, there are additional areas that require further examination to truly reap the full benefit and information that utilizing wearable and biosensors can provide. For example, work needs to be done to identify a minimal clinically relevant difference for physical activity variables derived from wearables and biosensors to know what is important from a patient perspective, as has been done in other areas relating to heart failure [18]. Furthermore, additional variables can be derived besides activity intensity and activity duration that may prove to be more informative about a specific patient population or allow for better discernment between groups of patients. Additionally, examination into association of physical activity variables and other clinical variables including but not limited to patient reported outcomes and physical limitation scores is of interest and would be necessary in hopes of utilizing wearables and biosensors further in clinical trials. Based on the work presented here and that of others [55], it suggests that data focusing on activity duration and activity intensity provide unique streams of information that can be utilized by algorithms to improve classification and predictive accuracy of the models. Further work is necessary to identify what are the most optimal derivations of activity duration and activity intensity, and if further sub-classifications and sub-derivations of these variables, would improve algorithmic accuracy. Additional working groups are currently examining devising an open-source approach for the analysis of wearable device data. Furthermore, as there are some differences dependent on the therapeutic area one is interested in collecting and analyzing wearable device data with, working groups such as the Heart Failure Collaboratory representing an innovative collaboration spanning the major pharmaceutical companies, biotechnological companies, and regulatory agencies like the FDA, are tasked with helping to address some of the regulatory hurdles implicit in these areas. While there clearly remain many open questions in the area of wearable devices and biosensors and their utility to clinical trials and more specifically heart failure, it is an exciting area of research that has the potential to change the landscape of clinical trials and improve patient lives for the better.

Bibliography

[1] Arun Bhatt. Evolution of clinical research: a history before and beyond james lind. *Perspectives in clinical research*, 1(1):6, 2010.

[2] Duolao Wang, Ameet Bakhai, and Nicola Maffulli. A primer for statistical analysis of clinical trials. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 19(8):874–881, 2003.

[3] Daniel C Malone, Lisa E Hines, and Jennifer S Graff. The good, the bad, and the different: a primer on aspects of heterogeneity of treatment effects. *Journal of Managed Care Pharmacy*, 20(6):555–563, 2014.

[4] John Kendall. Designing a research project: randomised controlled trials and their principles. *Emergency medicine journal: EMJ*, 20(2):164, 2003.

[5] Walter N Kernan, Catherine M Viscoli, Robert W Makuch, Lawrence M Brass, and Ralph I Horwitz. Stratified randomization for clinical trials. *Journal of clinical epidemiology*, 52(1):19–26, 1999.

[6] Annabelle S Slingerland, William H Herman, William K Redekop, Rob F Dijkstra, J Wouter Jukema, and Louis W Niessen. Stratified patient-centered care in type 2 diabetes: a cluster-randomized, controlled clinical trial of effectiveness and cost-effectiveness. *Diabetes Care*, 36(10):3054–3061, 2013.

[7] Justin A Ezekowitz, Christopher M O'Connor, Richard W Troughton, Wendimagegn G Alemayehu, Cynthia M Westerhout, Adriaan A Voors, Javed Butler, Carolyn SP Lam, Piotr Ponikowski, Michele Emdin, et al. N-terminal pro-b-type natriuretic peptide and clinical outcomes: vericiguat heart failure with reduced ejection fraction study. *Heart failure*, 8(11):931–939, 2020.

[8] John Greist, James Mundt, James Jefferson, and David Katzelnick. Comments on" why do clinical trials fail? the problem of measurement error in clinical trials: Time to test new paradigms?". *Journal of clinical psychopharmacology*, 27(5):535–537, 2007.

[9] G Savarese and LH Lund. Global public health burden of heart failure. card fail rev. 2017; 3 (1): 7-11, 2016.

[10] Peter A McCullough, Edward F Philbin, John A Spertus, Scott Kaatz, Keisha R Sandberg, and W Douglas Weaver. Confirmation of a heart failure epidemic: findings from the resource utilization among congestive heart failure (reach) study. *Journal of the American College of Cardiology*, 39(1):60–69, 2002.

[11] Simon Stewart, Kate MacIntyre, David J Hole, Simon Capewell, and John JV McMurray. More 'malignant'than cancer? five-year survival following a first admission for heart failure. *European journal of heart failure*, 3(3):315–322, 2001.

[12] Kirkwood F Adams, Jo Ann Lindenfeld, J Malcolm O Arnold, David W Baker, Denise H Barnard, Kenneth Lee Baughman, John P Boehmer, Prakash Deedwania, Sandra B Dunbar, Uri Elkayam, et al. Executive summary: Hfsa 2006 comprehensive heart failure practice guideline. *Journal of Cardiac Failure*, 12(1):10–38, 2006.

[13] Michael Urbich, Gary Globe, Krystallia Pantiri, Marieke Heisen, Craig Bennison, Heidi S Wirtz, and Gian Luca Di Tanna. A systematic review of medical costs associated with heart failure in the usa (2014–2020). *Pharmacoeconomics*, 38(11):1219–1236, 2020.

[14] Filip Machaj, Elżbieta Dembowska, Jakub Rosik, Bartosz Szostak, Małgorzata Mazurek-Mochol, and Andrzej Pawlik. New therapies for the treatment of heart failure: a summary of recent accomplishments. *Therapeutics and clinical risk management*, 15:147, 2019.

[15] Paul W Armstrong, Carolyn SP Lam, Kevin J Anstrom, Justin Ezekowitz, Adrian F Hernandez, Christopher M O'Connor, Burkert Pieske, Piotr Ponikowski, Sanjiv J Shah, Scott D Solomon, et al. Effect of vericiguat vs placebo on quality of life in patients with heart

failure and preserved ejection fraction: the vitality-hfpef randomized clinical trial. *Jama*, 324(15):1512–1521, 2020.

[16] Vanja Vlajnic, Chrysanthi Dori, Mercedeh Ghadessi, Maike Ahrens, Mathias Sachs, and Paolo Piraino. Wearable devices in clinical trials: Making an impact in the cardiovascular space., 2019.

[17] Irina Gaynanova, Naresh Punjabi, and Ciprian Crainiceanu. Modeling continuous glucose monitoring (cgm) data during sleep. *Biostatistics*, 23(1):223–239, 2022.

[18] Javed Butler, John A Spertus, Luke Bamber, Muhammad Shahzeb Khan, Lothar Roessig, Vanja Vlajnic, Josephine M Norquist, Kevin J Anstrom, Robert O Blaustein, Carolyn SP Lam, et al. Defining changes in physical limitation from the patient perspective: insights from the vitality-hfpef randomized trial. *European Journal of Heart Failure*, 2022.

[19] John A Spertus, Philip G Jones, Alexander T Sandhu, and Suzanne V Arnold. Interpreting the kansas city cardiomyopathy questionnaire in clinical trials and clinical care: Jacc state-of-the-art review. *Journal of the American College of Cardiology*, 76(20):2379–2390, 2020.

[20] Mona Fiuzat, Naomi Lowy, Norman Stockbridge, Marco Sbolli, Federica Latta, JoAnn Lindenfeld, Eldrin F Lewis, William T Abraham, John Teerlink, Mary Walsh, et al. Endpoints in heart failure drug development: history and future. *Heart Failure*, 8(6):429–440, 2020.

[21] Victoria H Stiles, Matthew Pearce, Isabel S Moore, Joss Langford, and Alex V Rowlands. Wrist-worn accelerometry for runners: objective quantification of training load. *Medicine and science in sports and exercise*, 50(11):2277, 2018.

[22] Haobo Li, Margaret H Hastings, James Rhee, Lena E Trager, Jason D Roh, and Anthony Rosenzweig. Targeting age-related pathways in heart failure. *Circulation research*, 126(4):533–551, 2020.

[23] Amytis Towfighi, Ling Zheng, and Bruce Ovbiagele. Sex-specific trends in midlife coronary heart disease risk and prevalence. *Archives of internal medicine*, 169(19):1762–1766, 2009.

[24] Luke C Cunningham, Gregg C Fonarow, Clyde W Yancy, Shubin Sheng, Roland A Matsouaka, Adam D DeVore, Hani Jneid, and Anita Deswal. Regional variations in heart failure quality and outcomes: Get with the guidelines–heart failure registry. *Journal of the American Heart Association*, 10(7):e018696, 2021.

[25] New York Heart Association, Criteria Committee, Martin Dolgin, et al. *Nomenclature and criteria for diagnosis of diseases of the heart and great vessels*. Little, Brown, 1994.

[26] Vanessa Blumer, Robert J Mentz, Jie-Lena Sun, Javed Butler, Marco Metra, Adriaan A Voors, Adrian F Hernandez, Christopher M O'Connor, and Stephen J Greene. Prognostic role of prior heart failure hospitalization among patients hospitalized for worsening chronic heart failure. *Circulation: Heart Failure*, 14(4):e007871, 2021.

[27] Shannon M Dunlay, Michael M Givertz, David Aguilar, Larry A Allen, Michael Chan, Akshay S Desai, Anita Deswal, Victoria Vaughan Dickson, Mikhail N Kosiborod, Carolyn L Lekavich, et al. Type 2 diabetes mellitus and heart failure: a scientific statement from the american heart association and the heart failure society of america: this statement does not represent an update of the 2017 acc/aha/hfsa heart failure guideline update. *Circulation*, 140(7):e294–e324, 2019.

[28] Johannes Brachmann, Christian Sohns, Dietrich Andresen, Jürgen Siebels, Susanne Sehner, Luca Boersma, Béla Merkely, Evgeny Pokushalov, Prashanthan Sanders, Heribert Schunkert, et al. Atrial fibrillation burden and clinical outcomes in heart failure: the castle-af trial. *Clinical Electrophysiology*, 7(5):594–603, 2021.

[29] Xinghe Huang, Jiamin Liu, Lihua Zhang, Bin Wang, Xueke Bai, Shuang Hu, Fengyu Miao, Aoxi Tian, Tingxuan Yang, Yan Li, et al. Systolic blood pressure and one-year clinical outcomes in patients hospitalized for heart failure. *Frontiers in cardiovascular medicine*, page 885, 2022.

[30] David H Smith, Micah L Thorp, Jerry H Gurwitz, David D McManus, Robert J Goldberg, Larry A Allen, Grace Hsu, Sue Hee Sung, David J Magid, and Alan S Go. Chronic kidney disease and outcomes in heart failure with preserved versus reduced ejection fraction: the cardiovascular research network preserve study. *Circulation: Cardiovascular Quality and Outcomes*, 6(3):333–342, 2013.

[31] Zhuo Chen, Qian Lin, Jingen Li, Xinyi Wang, Jianqing Ju, Hao Xu, and Dazhuo Shi. Estimated glomerular filtration rate is associated with an increased risk of death in heart failure patients with preserved ejection fraction. *Frontiers in Cardiovascular Medicine*, 8:325, 2021.

[32] Satish Kenchaiah, Stuart J Pocock, Duolao Wang, Peter V Finn, Leonardo AM Zornoff, Hicham Skali, Marc A Pfeffer, Salim Yusuf, Karl Swedberg, Eric L Michelson, et al. Body mass index and prognosis in patients with chronic heart failure: insights from the candesartan in heart failure: Assessment of reduction in mortality and morbidity (charm) program. *Circulation*, 116(6):627–636, 2007.

[33] Faiez Zannad, Angeles Alonso Garcia, Stefan D Anker, Paul W Armstrong, Gonzalo Calvo, John GF Cleland, Jay N Cohn, Kenneth Dickstein, Michael J Domanski, Inger Ekman, et al. Clinical outcome endpoints in heart failure trials: a european society of cardiology heart failure association consensus document. *European journal of heart failure*, 15(10):1082–1094, 2013.

[34] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61:1–36, 2014.

[35] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[36] Alexander Topchy, Anil K Jain, and William Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1866–1881, 2005.

[37] Christopher R John, David Watson, Dominic Russ, Katriona Goldmann, Michael Ehrenstein, Costantino Pitzalis, Myles Lewis, and Michael Barnes. M3c: Monte carlo reference-based consensus clustering. *Scientific reports*, 10(1):1–14, 2020.

[38] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[39] David R Cox. Regression models and life tables (with discussion). *JR Stat Soc*, 34:187–220, 1972.

[40] Mike J Bradburn, Taane G Clark, Sharon B Love, and Douglas G Altman. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431–436, 2003.

[41] JF Lawless and Kishore Singhal. Efficient screening of nonnormal regression models. *Biometrics*, pages 318–327, 1978.

[42] Frank E Harrell Jr, Maintainer Frank E Harrell Jr, and Depends Hmisc. Package 'rms'. *Vanderbilt University*, 229, 2017.

[43] Terry M Therneau and Patricia M Grambsch. The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer, 2000.

[44] Jindrich Spinar, Lenka Spinarova, Filip Malek, Ondrej Ludka, Jan Krejci, Petr Ostadal, Dagmar Vondrakova, Karel Labr, Monika Spinarova, Monika Pavkova Goldbergova, et al. Prognostic value of nt-probnp added to clinical parameters to predict two-year prognosis of chronic heart failure patients with mid-range and reduced ejection fraction–a report from far nhl prospective registry. *PloS one*, 14(3):e0214363, 2019.

[45] Sepideh Sokhanvar, Mahdiye Shekhi, Saeedeh Mazlomzadeh, and Zahra Golmohammadi. The relationship between serum nt–pro-bnp levels and prognosis in patients with systolic heart failure. *Journal of cardiovascular and thoracic research*, 3(2):57, 2011.

[46] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[47] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

[48] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[49] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):1–11, 2008.

[50] Hemant Ishwaran, Udaya B Kogalur, Eiran Z Gorodeski, Andy J Minn, and Michael S Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010.

[51] Hemant Ishwaran, Udaya B Kogalur, and Maintainer Udaya B Kogalur. Package 'randomforestsrc'. *breast*, 6:1, 2022.

[52] Philipp Probst. *Hyperparameters, tuning and meta-learning for random forest and other machine learning algorithms*. PhD thesis, lmu, 2019.

[53] John Ehrlinger. ggrandomforests: Exploring random forest survival. *arXiv preprint arXiv:1612.08974*, 2016.

[54] Pew Research Center. Mobile devices fact sheet. https://www.pewresearch.org/internet/fact-sheet/mobile/, 2021.

[55] Bunny J Pozehl, Rita Mcguire, Kathleen Duncan, Melody Hertzog, Pallav Deka, Joseph Norman, Nancy T Artinian, Matthew A Saval, and Steven J Keteyian. Accelerometer-measured

daily activity levels and related factors in patients with heart failure. *The Journal of cardiovascular nursing*, 33(4):329, 2018.

Appendix A

Programming Code

```
  data work.advsact;
   set ads.advsact;
   if AVISITN = 1 and PARAMN=11 and VSEVINTX="INTRADAY"
   and RANDFL="Y" and DTYPE ^= "AVERAGE" and ADT ^= .;
 run;

proc summary data=work.advsact nway;
    class USUBJID ADT;
    var AVAL;
    output out=advsact_means mean=mean;
run;

proc sql;
create table actdaycount as
     select USUBJID
     , ADT
     ,count(*) as Count
     from advsact_means
group by USUBJID
   ;
quit;
```

```
data work.actday_cleaned;
   set work.actdaycount;
   if Count GE 7;
RUN;

proc sort data=work.actday_cleaned;
   by USUBJID ADT;
RUN;

proc sort data=work.advsact_means;
   by USUBJID ADT;
RUN;

data work.act_merge;
   merge work.actday_cleaned(in=a) work.advsact_means(in=b);
   by USUBJID ADT;
   if a and b;
RUN;

data work.act_merge_ct7;
   set work.act_merge;
   if Count=7;
RUN;

data work.act_merge_ct8;
   set work.act_merge;
   if Count=8;
```

```
RUN;

proc sort data=work.act_merge_ct8;
   by USUBJID _FREQ_;
RUN;

data work.act_merge_ct8_clean;
   set work.act_merge_ct8;
   if first.USUBJID then delete;
   by USUBJID;
RUN;

proc sql;
create table work.act_merge_ct8_final as
     select USUBJID
     , ADT
     ,count(*) as Count_v2
     from work.act_merge_ct8_clean
group by USUBJID
   ;
quit;

data work.act_merge_ct9;
   set work.act_merge;
   if Count=9;
RUN;
```

```
data work.act_merge_ct9;
   set work.act_merge;
   if Count=9;
RUN;

proc sort data=work.act_merge_ct9;
   by USUBJID _FREQ_;
RUN;

data work.act_merge_ct9_clean_step1;
   set work.act_merge_ct9;
   if first.USUBJID then delete;
   by USUBJID;
RUN;

data work.act_merge_ct9_clean_step2;
   set work.act_merge_ct9_clean_step1;
   if first.USUBJID then delete;
   by USUBJID;
RUN;

proc sql;
create table work.act_merge_ct9_final as
      select USUBJID
      , ADT
      ,count(*) as Count_v2
      from work.act_merge_ct9_clean_step2
```

```
group by USUBJID
   ;
quit;

proc sort data=work.actday_cleaned;
   by USUBJID ADT;
RUN;

proc sort data=work.advsact_means;
   by USUBJID ADT;
RUN;

proc sort data=work.act_merge_ct7;
   by USUBJID ADT;
RUN;

proc sort data=work.act_merge_ct8_final;
   by USUBJID ADT;
RUN;

proc sort data=work.act_merge_ct9_final;
   by USUBJID ADT;
RUN;

data work.act_combined;
   merge work.act_merge_ct7 work.act_merge_ct8_final
   work.act_merge_ct9_final;
   by USUBJID ADT;
RUN;

data work.act_combined_dropmeans;
   set work.act_combined;
   drop mean;
RUN;

proc sort data=work.act_combined_dropmeans;
   by USUBJID ADT;
RUN;

data work.act_final;
   merge work.act_combined_dropmeans(in=a)
   work.advsact_means(in=b);
   by USUBJID ADT;
   if a and b;
RUN;

data work.act_final;
   set work.act_final;
   drop Count Count_v2 _FREQ_ _TYPE_;
RUN;

data work.advsint;
   set ads.advsint;
   if AVISITN=1 and PARAMN=8 and RANDFL="Y"
   and DTYPE ^= "AVERAGE" and ADT ^= .;
run;

proc summary data=work.advsint nway;
    class USUBJID ADT;
    var AVAL;
    output out=advsint_means mean=mean;
run;

proc sql;
create table actintcount as
     select USUBJID
     , ADT
     ,count(*) as Count
     from advsint_means
group by USUBJID
   ;
quit;

data work.actint_cleaned;
   set work.actintcount;
   if Count GE 7;
RUN;

proc sort data=work.actint_cleaned;
   by USUBJID ADT;
RUN;
```

```
proc sort data=work.advsint_means;
   by USUBJID ADT;
RUN;

data work.int_merge;
   merge work.actint_cleaned(in=a)
   work.advsint_means(in=b);
   by USUBJID ADT;
   if a and b;
RUN;

data work.int_merge_ct7;
   set work.int_merge;
   if Count=7;
RUN;

data work.int_merge_ct8;
   set work.int_merge;
   if Count=8;
RUN;

proc sort data=work.int_merge_ct8;
   by USUBJID _FREQ_;
RUN;

data work.int_merge_ct8_clean;
```

```
   set work.int_merge_ct8;
   if first.USUBJID then delete;
   by USUBJID;
RUN;

proc sql;
create table work.int_merge_ct8_final as
     select USUBJID
     , ADT
     ,count(*) as Count_v2
     from work.int_merge_ct8_clean
group by USUBJID
   ;
quit;

data work.int_merge_ct9;
   set work.int_merge;
   if Count=9;
RUN;

data work.int_merge_ct9;
   set work.int_merge;
   if Count=9;
RUN;

proc sort data=work.int_merge_ct9;
   by USUBJID _FREQ_;
```

```
RUN;

data work.int_merge_ct9_clean_step1;
   set work.int_merge_ct9;
   if first.USUBJID then delete;
   by USUBJID;
RUN;

data work.int_merge_ct9_clean_step2;
   set work.int_merge_ct9_clean_step1;
   if first.USUBJID then delete;
   by USUBJID;
RUN;

proc sql;
create table work.int_merge_ct9_final as
     select USUBJID
     , ADT
     ,count(*) as Count_v2
     from work.int_merge_ct9_clean_step2
group by USUBJID
   ;
quit;

proc sort data=work.actint_cleaned;
   by USUBJID ADT;
RUN;
```

```
proc sort data=work.advsint_means;
   by USUBJID ADT;
RUN;

proc sort data=work.int_merge_ct7;
   by USUBJID ADT;
RUN;

proc sort data=work.int_merge_ct8_final;
   by USUBJID ADT;
RUN;

proc sort data=work.int_merge_ct9_final;
   by USUBJID ADT;
RUN;

data work.int_combined;
   merge work.int_merge_ct7 work.int_merge_ct8_final
   work.int_merge_ct9_final;
   by USUBJID ADT;
RUN;

data work.int_combined_dropmeans;
   set work.int_combined;
   drop mean;
RUN;
```

```
proc sort data=work.int_combined_dropmeans;
   by USUBJID ADT;
RUN;

data work.int_final;
   merge work.int_combined_dropmeans(in=a)
   work.advsint_means(in=b);
   by USUBJID ADT;
   if a and b;
RUN;

data work.int_final;
   set work.int_final;
   drop Count Count_v2 _FREQ_ _TYPE_;
RUN;

data work.int_final;
   set work.int_final;
   rename mean = mean_INT;
RUN;

proc sort data=work.int_final;
   by USUBJID ADT;
RUN;

data work.act_final;
```

```
   set work.act_final;
   rename mean = mean_ACT;
RUN;

proc sort data=work.act_final;
   by USUBJID ADT;
RUN;

data work.act_int_merge;
   merge work.int_final(in=a) work.act_final(in=b);
   by USUBJID ADT;
   if a and b;
RUN;

proc sql;
create table merge_counts as
     select USUBJID
     , ADT
     ,count(*) as Count
     from work.act_int_merge
group by USUBJID
   ;
quit;

data work.act_int_merge_comp;
   merge work.int_final(in=a) work.act_final(in=b);
   by USUBJID ADT;
```

```
RUN;

proc sql;
create table count_comp as
     select USUBJID
     , ADT
     ,count(*) as Count
     from work.act_int_merge_comp
group by USUBJID
   ;
quit;

proc sort data=count_comp;
   by USUBJID ADT;
RUN;

proc sort data=work.act_int_merge_comp;
by USUBJID ADT;
RUN;

data work.merge_comp_counts;
  merge work.act_int_merge_comp count_comp;
   by USUBJID ADT;
RUN;

data work.merge_comp_missing;
set work.merge_comp_counts;
```

```
array vars(3) mean_INT mean_ACT;
numMissing = cmiss(of vars[*]);
run;

data work.merge_comp_missing_ct8;
   set work.merge_comp_missing;
   if Count=8;
RUN;

proc sort data=work.merge_comp_missing_ct8;
   by USUBJID numMissing;
RUN;

data work.merge_comp_missing_ct8_cln;
   set work.merge_comp_missing_ct8;
   if last.USUBJID then delete;
   by USUBJID;
RUN;

data work.act_intraday_avg;
   set ads.advsact;
   if AVISITN=1 and VSEVINTX="INTRADAY"
   and RANDFL="Y" and DTYPE="AVERAGE";
   keep USUBJID AVAL;
   rename AVAL=act_avg;
RUN;
```

```
data work.int_intraday_avg;
   set ads.advsint;
   if AVISITN=1 and RANDFL="Y" and ABLFL="Y"
   and PARAMN=8;
   keep USUBJID BASE;
   rename BASE=int_avg;
RUN;

proc sort data = work.act_intraday_avg;
   by USUBJID;
RUN;

proc sort data = work.int_intraday_avg;
   by USUBJID;
RUN;

data work.merge_averages;
  merge work.act_intraday_avg(in=a)
  work.int_intraday_avg(in=b);
  by USUBJID;
  if a and b;
RUN;

proc sort data=work.merge_comp_counts;
   by USUBJID;
RUN;
```

```
proc sort data=work.merge_averages;
   by USUBJID;
RUN;

data work.merge_comp_avgs;
   merge work.merge_comp_counts(in=a)
   work.merge_averages(in=b);
   by USUBJID;
RUN;

data work.merge_comp_avgs_ct7;
   set work.merge_comp_avgs;
   if Count=7;
RUN;

data work.comp_avg_counter;
   set work.merge_comp_avgs_ct7;
   by USUBJID;
   if first.USUBJID then COUNTER_act=0;
   COUNTER_act+(mean_ACT eq .);
   if first.USUBJID then COUNTER_int=0;
   COUNTER_int+(mean_INT eq .);
RUN;

data work.comp_avg_counter_identify;
   set work.comp_avg_counter;
   by USUBJID;
   if COUNTER_act > 1 or COUNTER_int > 1
   then remove=1;
   else remove=0;
RUN;

proc sql;
   create table work.pat_remove as
   select USUBJID, sum(remove) as remove_sum
   from work.comp_avg_counter_identify
   group by USUBJID;
QUIT;

proc sort data=work.comp_avg_counter_identify;
   by USUBJID;
RUN;

proc sort data=work.pat_remove;
   by USUBJID;
RUN;

data work.comp_avg_merged_c7_prep;
   merge work.comp_avg_counter_identify
   work.pat_remove;
   by USUBJID;
RUN;

data work.comp_avg_merged_c7_clean;
```

```
   set work.comp_avg_merged_c7_prep;
   if remove_sum = 0;
RUN;

data work.merge_comp_avgs_8;
   set work.merge_comp_avgs;
   if Count = 8;
RUN;

proc sort data=work.merge_comp_avgs_8;
   by USUBJID ADT;
RUN;

proc sort data=work.merge_comp_missing_ct8_cln;
   by USUBJID ADT;
RUN;

data work.merge_clean_ct8_part1;
   merge work.merge_comp_avgs_8(in=a)
   work.merge_comp_missing_ct8_cln(in=b);
   by USUBJID ADT;
   if a and b;
RUN;

data work.merge_clean_ct8_part1_imp;
   set work.merge_clean_ct8_part1;
   if mean_ACT = . then mean_ACT = act_avg;
   if mean_INT = . then mean_INT = int_avg;
RUN;

proc sort data=work.merge_clean_ct8_part1_imp;
   by USUBJID;
RUN;

proc sort data=work.comp_avg_merged_c7_clean;
   by USUBJID;
RUN;

data work.final_intraday_part1;
   merge work.merge_clean_ct8_part1_imp
   work.comp_avg_merged_c7_clean;
   by USUBJID;
   keep USUBJID ADT mean_INT mean_ACT;
RUN;



proc sgplot data=work.final_intraday_part1;
   histogram mean_ACT;
RUN;

proc sgplot data=work.final_intraday_part1;
   histogram mean_INT;
RUN;
```

```
data work.final_intraday_part1;
   set work.final_intraday_part1;
   logmean_ACT = log(mean_ACT);
   logmean_INT = log(mean_INT);
RUN;

proc sgplot data=work.final_intraday_part1;
   histogram logmean_ACT;
RUN;

proc sgplot data=work.final_intraday_part1;
   histogram logmean_INT;
RUN;

data work.final_intraday_part2_ct;
   set work.final_intraday_part1;
   count + 1;
   by USUBJID;
   if first.USUBJID then count=1;
RUN;

data work.final_intraday_part2_rn_per1;
   set work.final_intraday_part2_ct;
   rename mean_INT = mean_INT_per1;
   rename mean_ACT = mean_ACT_per1;
   drop logmean_INT logmean_ACT count ADT;
RUN;
```

```
data work.final_intraday_part2_ct_dt;
   length date $5;
   set work.final_intraday_part2_ct;
   date = cats(’date’,count);
   drop count ADT;
RUN;

proc sql;
   create table work.avg_per_pat_INT as
   select USUBJID, mean(mean_INT) as avg_mean_INT
   from work.final_intraday_part2_ct_dt
   group by USUBJID;
QUIT;

proc sql;
   create table work.avg_per_pat_ACT as
   select USUBJID, mean(mean_ACT) as avg_mean_ACT
   from work.final_intraday_part2_ct_dt
   group by USUBJID;
QUIT;

data avg_pat_ACT_INT;
   merge work.avg_per_pat_INT work.avg_per_pat_ACT;
   by USUBJID;
run:
```

```
proc transpose data=work.final_intraday_part2_ct_dt
out=wide_ACT prefix=per1_ACT;
   by USUBJID;
   ID date;
   var mean_ACT;
RUN;

proc transpose data=work.final_intraday_part2_ct_dt
out=wide_INT prefix=per1_INT;
   by USUBJID;
   ID date;
   var mean_INT;
RUN;

proc transpose data=work.final_intraday_part2_ct_dt
out=wide_logACT prefix=per1_logACT;
   by USUBJID;
   ID date;
   var logmean_ACT;
RUN;

proc transpose data=work.final_intraday_part2_ct_dt
out=wide_logINT prefix=per1_logINT;
   by USUBJID;
   ID date;
   var logmean_INT;
RUN;
```

```
data work.wide_final;
   merge wide_ACT(drop=_name_) wide_INT(drop=_name_)
   wide_logACT(drop=_name_) wide_logINT(drop=_name_);
   by USUBJID;
   drop _LABEL_;
run:

proc standard data=wide_final mean=0 STD=1
out=wide_final_std;
   var per1_ACTdate1--per1_INTdate7;
RUN;

proc standard data=wide_final mean=0 STD=1
out=wide_logfinal_std;
   var per1_logACTdate1--per1_logINTdate7;
RUN;

data work.advsinmn;
   set ads.advsinmn;
run;

data work.adsvinmn_subj;
   set work.advsinmn;
   by USUBJID;
   if first.USUBJID then output;
RUN;
```

```
libname results "~/results";

data work.adtte;
   set ads.adtte;
RUN;

proc copy in=work out=results;
   select adtte;
RUN;

proc copy in=work out=results;
   select wide_final;
RUN;

proc copy in=work out=results;
   select wide_final_std;
RUN;

proc copy in=work out=results;
   select adsvinmn_subj;
RUN;

proc copy in=work out=results;
   select final_intraday_part2_ct_dt;
RUN;
```

```
proc copy in=work out=results;
   select avg_pat_ACT_INT;
RUN;
```

K-means clustering

```
rm(list = ls())

library(namespace)
library(rlang)
library(data.table)
library(tidyverse)
library(stringr)
library(readr)
library(haven)
library(lubridate)
library(dplyr)
library(cluster)
library(factoextra)
library(formattable)
library(NbClust)
library(psych)

wide_final <- read_sas(data_file = /wide_final.sas7bdat’)
wide_final_std␣<-␣read_sas(data_file␣=␣/wide_final_std.sas7bdat’)
wide_final_std_cln <- select(wide_final_std, -USUBJID)
```

```
wide_final_cln <- select(wide_final_std, -USUBJID)
advsinmx <- read_sas(data_file = ’/adsvinmn_subj.sas7bdat’)
long_final <- read_sas(data_file = ’/final_intraday_part2_ct_dt.
   sas7bdat’)

df <- wide_final_std_cln
k3 <- kmeans(wide_final_std_cln, centers=3, nstart=20)

set.seed(123)

wss <- function(k) {
 kmeans(wide_final_std_cln, k, nstart = 10 )$tot.withinss
}

k.values <- 1:15

wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values,
    type="b", pch = 19, frame = FALSE,
    xlab="Number␣of␣clusters␣K",
    ylab="Total␣within-clusters␣sum␣of␣squares")

set.seed(123)
fviz.nbclust(wide_final_std_cln, kmeans, method="wss")

avg_sil <- function(k) {
  km.res <- kmeans(df, centers = k, nstart = 25)
  ss <- silhouette(km.res$cluster, dist(df))
  mean(ss[, 3])
}

k.values <- 2:15

avg_sil_values <- map_dbl(k.values, avg_sil)

plot(k.values, avg_sil_values,
    type = "b", pch = 19, frame = FALSE,
    xlab = "Number␣of␣clusters␣K",
    ylab = "Average␣Silhouettes")

set.seed(123)
gap_stat <- clusGap(df, FUN = kmeans, nstart = 25,
               K.max = 20, B = 50)
print(gap_stat, method = "firstmax")
fviz_gap_stat(gap_stat)

df <-scale(wide_final_cln)
distance <- get_dist(df)
fviz_dist(distance, gradient = list(low = "#00AFBB",
mid = "white", high = "#FC4E07"))

k2 <- kmeans(df, centers = 2, nstart = 25)
str(k2)
```

```
k2
fviz_cluster(k2, data = df)

k3 <- kmeans(df, centers = 3, nstart = 25)
str(k3)
k3
fviz_cluster(k3, data = df)

k4 <- kmeans(df, centers = 4, nstart = 25)
str(k4)
k4
fviz_cluster(k4, data = df)

set.seed(123)
fviz_nbclust(df, kmeans, method = "wss")

set.seed(123)
fviz_nbclust(df, kmeans, method = "silhouette")

set.seed(123)
fviz_nbclust(df, kmeans, nstart=25,
method="gap_stat", nboot=50)

set.seed(123)
nb <- NbClust(df, distance = "euclidean",
      min.nc = 2, max.nc = 15,method="kmeans")
fviz_nbclust(nb)
```

```
print(gap_stat, method = "firstmax")
fviz_gap_stat(gap_stat)

USUBJID <- wide_final$USUBJID


id_cluster2 <- cbind(USUBJID, k2$cluster)
id_cluster3 <- cbind(id_cluster2, k3$cluster)
final_clusters <- merge(id_cluster3, advsinmx, by="USUBJID")
df_long_3 <- merge(id_cluster3, long_final, by="USUBJID")
final_clusters_step1 <- final_clusters[-c(4:118)]
df_3 <- final_clusters_step1

df$AGEGR01 = as.factor(df$AGEGR01)
df$age = as.factor(df$age)
split(df, df$cluster) |> summary()

tapply(df$AGEGR01, df$V2, function(x) table(x)/length(x))
tapply(df$age, df$cluster, function(x) table(x)/length(x))
tapply(df$O2, df$cluster, function(x) table(x)/length(x))

table(df_3$V3)

fviz_cluster(k3, data = df)
df_3$AGEGR01 = as.factor(df_3$AGEGR01)
df_3$AGEGR02 = as.factor(df_3$AGEGR02)
df_3$V3 = as.factor(df_3$V3)
```

```
tapply(df_3$AGEGR01, df_3$V3, function(x) table(x)/length(x))
tapply(df_3$AGEGR02, df_3$V3, function(x) table(x)/length(x))

df_long_3$V3 = as.factor(df_long_3$V3)

describeBy(df_long_3$mean_INT, group=df_long_3$V3, mat=TRUE,
   digits=3)
describeBy(df_long_3$mean_ACT, group=df_long_3$V3, mat=TRUE,
   digits=3)

df_3$SEX = as.factor(df_3$SEX)
table(df_3$V3, df_3$SEX)

tapply(df_3$SEX, df_3$V3, function(x) table(x)/length(x))

df_3$RACE = as.factor(df_3$RACE)
table(df_3$V3, df_3$RACE)

df_3$CNTYGR1N = as.factor(df_3$CNTYGR1N)
table(df_3$V3, df_3$CNTYGR1N)

tapply(df_3$CNTYGR1N, df_3$V3, function(x) table(x)/length(x))

df_3$NTBNP = as.factor(df_3$NTBNP)
tapply(df_3$NTBNP, df_3$V3, function(x) table(x)/length(x))
```

```
df_3$NYHAFUC = as.factor(df_3$NYHAFUC)
tapply(df_3$NYHAFUC, df_3$V3, function(x) table(x)/length(x))

df_3$BETABLK = as.factor(df_3$BETABLK)
tapply(df_3$BETABLK, df_3$V3, function(x) table(x)/length(x))

df_3$PRIHOSP = as.factor(df_3$PRIHOSP)
tapply(df_3$PRIHOSP, df_3$V3, function(x) table(x)/length(x))

df_3$DIABETE = as.factor(df_3$DIABETE)
tapply(df_3$DIABETE, df_3$V3, function(x) table(x)/length(x))

df_3$ATRFIBR = as.factor(df_3$ATRFIBR)
tapply(df_3$ATRFIBR, df_3$V3, function(x) table(x)/length(x))

df_3$HYPERTE = as.factor(df_3$HYPERTE)
tapply(df_3$HYPERTE, df_3$V3, function(x) table(x)/length(x))

df_3$MAXACE = as.factor(df_3$MAXACE)
tapply(df_3$MAXACE, df_3$V3, function(x) table(x)/length(x))

df_3$MAXARB = as.factor(df_3$MAXARB)
tapply(df_3$MAXARB, df_3$V3, function(x) table(x)/length(x))

df_3$BASEGFR = as.factor(df_3$BASEGFR)
tapply(df_3$BASEGFR, df_3$V3, function(x) table(x)/length(x))
```

```
df_3$BMIGR1 = as.factor(df_3$BMIGR1)
tapply(df_3$BMIGR1, df_3$V3, function(x) table(x)/length(x))

df_3$BMIGR2 = as.factor(df_3$BMIGR2)
tapply(df_3$BMIGR2, df_3$V3, function(x) table(x)/length(x))

df_3$LVEFB25 = as.factor(df_3$LVEFB25)
tapply(df_3$LVEFB25, df_3$V3, function(x) table(x)/length(x))

df_3$LVEFHIM = as.factor(df_3$LVEFHIM)
tapply(df_3$LVEFHIM, df_3$V3, function(x) table(x)/length(x))

describeBy(df_3$CNTYGR1N, group=df_3$V3, mat=TRUE, digits=3)
tapply(df_long_3$mean_INT, df_long_3$V3, min, mean, median,
   maximum)

descriptive_intensity <- df_long_3 %>%
  group_by(V3) %>%
  summarize(mean_INT)
summarize(mean_INT~V3, data=df_long_3, digits=3)

  final_clusters_step1 %>%
  group_by(V2) %>%
  summarise(PRIHOSPN = n())%>%
  mutate(freq = formattable::percent(PRIHOSPN / sum(PRIHOSPN)))
    %>%
  arrange(desc(freq))
```

```
split(df_long_3, df_long_3$V3) | summary()


Consensus clustering

rm(list = ls())


library(ellipsis)

library(namespace)

library(rlang)

library(data.table)

library(tidyverse)

library(stringr)

library(readr)

library(haven)

library(lubridate)

library(tidyverse)

library(cluster)

library(factoextra)

library(formattable)

library(NbClust)

library(psych)

library(xtable)

library(table1)

library(survival)

library(survminer)

library(survivalAnalysis)


wide_final<- read_sas(data_file = ’/wide_final.sas7bdat’)
```

```
wide_final_std <- read_sas(data_file = '/wide_final_std.sas7bdat'
   )
wide_final_std_cln <- select(wide_final_std, -USUBJID)
wide_final_cln <- select(wide_final_std, -USUBJID)
advsinmx <- read_sas(data_file = '/adsvinmn_subj.sas7bdat')

long_final <- read_sas(data_file = '/final_intraday_part2_ct_dt.
   sas7bdat')
avg_pat <- read_sas(data_file = '/avg_pat_ACT_INT.sas7bdat')
long_final_comb <- read_sas(data_file = '/final_intraday_comb.
   sas7bdat')

wide_final_complete <-
read_sas(data_file = '/wide_final_log_comb.sas7bdat')
wide_final_complete_cln <- select(wide_final_complete, -USUBJID)
advsinmx <- read_sas(data_file = '/adsvinmn_subj.sas7bdat')

library(M3C)
df <- wide_final_complete_cln
df_t <- t(df)
df_t <- as.data.frame(df_t)
colnames(df_t) <- cbind(1,1:109)
colnames(df_t) <- paste0("foo_", colnames(df_t))

df_test <- df_t
df_test2 <- df_test %>%
  set_names(c(seq_along(df_test)))
```

```
colnames(df_test2) <- paste0("foo_", colnames(df_test2))
df_t <- df_test2


test_M3C <- M3C(df)
res_t_km <- M3C(df_t, removeplots = FALSE, iters=25,
objective=’entropy’, fsize=8, lthick=1, dotsize=1.25,
clusteralg=c("pam"))


res_t_km$plots[[1]]
res_t_km$plots[[2]]
res_t_km$plots[[3]]
res_t_km$plots[[4]]



annon_t_km <- res_t_km$realdataresults[[4]]$ordered_annotation
annon_t_km <- as.data.frame(annon_t_km)
annon_t_km$consensuscluster <- as.factor(annon_t_km$
   consensuscluster)
table(annon_t_km$consensuscluster)


df_update_km <- tibble::rownames_to_column(annon_t_km, "patinfo")
df_update_km$patinfo <- gsub("foo_V","",as.character(df_update_km
   $patinfo))
df_update_km$patinfo <- as.numeric(df_update_km$patinfo)
df_update_km <-as.data.frame(df_update_km)


df_update_km_sort <- df_update_km[order(df_update_km$patinfo),]
```

```
wide_final_complete_sort <-
wide_final_complete[order(wide_final_complete$USUBJID),]

intermediate_merge <- cbind(wide_final_complete_sort, df_update_
   km_sort)
names(intermediate_merge)[names(intermediate_merge)
== "consensuscluster"] <- "Cluster"

final_clinical_consensus_cluster <-
merge(intermediate_merge, advsinmx, by="USUBJID")

names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster)
== "consensuscluster"] <- "Cluster"
final_clinical_consensus_cluster$Cluster <- as.factor(final_
   clinical_consensus_cluster$Cluster)

final_clinical_consensus_cluster$CNTYGR1N <- as.character(final_
   clinical_consensus_cluster$CNTYGR1N)

final_clinical_consensus_cluster$
CNTYGR1N[final_clinical_consensus_cluster$CNTYGR1N == "5"]
<- "Eastern␣Europe"
final_clinical_consensus_cluster$
CNTYGR1N[final_clinical_consensus_cluster$CNTYGR1N == "6"]
<- "Western␣Europe␣and␣Israel"
final_clinical_consensus_cluster$
```

```
CNTYGR1N[final_clinical_consensus_cluster$CNTYGR1N == "8"]
<- "North␣America"

final_clinical_consensus_cluster$BASEGFR <- as.character(final_
   clinical_consensus_cluster$BASEGFR)
final_clinical_consensus_cluster$
BASEGFR[final_clinical_consensus_cluster$BASEGFR == ""]
<- "Missing"

final_clinical_consensus_cluster$LVEFHIM <- as.character(final_
   clinical_consensus_cluster$LVEFHIM)
final_clinical_consensus_cluster$
LVEFHIM[final_clinical_consensus_cluster$LVEFHIM == ""]
<- "Missing"

final_clinical_consensus_cluster$AGEGR01 = as.factor(final_
   clinical_consensus_cluster$AGEGR01)
final_clinical_consensus_cluster$AGEGR02 = as.factor(final_
   clinical_consensus_cluster$AGEGR02)
final_clinical_consensus_cluster$SEX =
as.factor(final_clinical_consensus_cluster$SEX)
final_clinical_consensus_cluster$RACE =
as.factor(final_clinical_consensus_cluster$RACE)
final_clinical_consensus_cluster$CNTYGR1N = as.factor(final_
   clinical_consensus_cluster$CNTYGR1N)
final_clinical_consensus_cluster$NTBNP = as.factor(final_clinical
   _consensus_cluster$NTBNP)
```

```
final_clinical_consensus_cluster$NYHAFUC = as.factor(final_
   clinical_consensus_cluster$NYHAFUC)
final_clinical_consensus_cluster$BETABLK = as.factor(final_
   clinical_consensus_cluster$BETABLK)
final_clinical_consensus_cluster$PRIHOSP = as.factor(final_
   clinical_consensus_cluster$PRIHOSP)
final_clinical_consensus_cluster$DIABETE = as.factor(final_
   clinical_consensus_cluster$DIABETE)

final_clinical_consensus_cluster$ATRFIBR = as.factor(final_
   clinical_consensus_cluster$ATRFIBR)
final_clinical_consensus_cluster$HYPERTE = as.factor(final_
   clinical_consensus_cluster$HYPERTE)
final_clinical_consensus_cluster$MAXACE = as.factor(final_
   clinical_consensus_cluster$MAXACE)
final_clinical_consensus_cluster$MAXARB = as.factor(final_
   clinical_consensus_cluster$MAXARB)

final_clinical_consensus_cluster$BASEGFR = as.factor(final_
   clinical_consensus_cluster$BASEGFR)
final_clinical_consensus_cluster$BMIGR1 = as.factor(final_
   clinical_consensus_cluster$BMIGR1)
final_clinical_consensus_cluster$LVEFB25 = as.factor(final_
   clinical_consensus_cluster$LVEFB25)
final_clinical_consensus_cluster$LVEFHIM = as.factor(final_
   clinical_consensus_cluster$LVEFHIM)
```

```
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "AGEGR01"]
<- "Age_Group"
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "CNTYGR1N"]
<- "Country_Group"
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "NYHAFUC"]
<- "NYHA"
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "PRIHOSP"]
<- "Prior_HF_hosp"
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "DIABETE"]
<- "Diabetes"
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "ATRFIBR"]
<- "Afib"
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "HYPERTE"]
<- "Hypertension"
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "BASEGFR"]
<- "EGFR"
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "LVEFHIM"]
<- "LVEF"
```

```
names(final_clinical_consensus_cluster)
[names(final_clinical_consensus_cluster) == "BMIGR1"]
<- "BMI"
full4_table_test <- table1(~ factor(Age_Group)
+ factor(SEX) + factor(Country_Group) + factor(NTBNP)
+ factor(NYHA) + factor(Prior_HF_hosp) + factor(Diabetes)
+ factor(Afib) + factor(Hypertension) + factor(EGFR)
+ factor(BMI) | Cluster, data=final_clinical_consensus_cluster)
full4_table_test

long_final_cluster <-
merge(long_final_comb, USUB_cluster, by="USUBJID")

names(long_final_cluster)[names(long_final_cluster) ==
"mean_INT_per1"] <- "Act_Int_Per1"
names(long_final_cluster)[names(long_final_cluster) ==
"mean_INT_per2"] <- "Act_Int_Per2"
names(long_final_cluster)[names(long_final_cluster) ==
"mean_INT_per3"] <- "Act_Int_Per3"
names(long_final_cluster)[names(long_final_cluster) ==
"mean_INT_per4"] <- "Act_Int_Per4"

names(long_final_cluster)[names(long_final_cluster) ==
"mean_ACT_per1"] <- "Act_Dur_Per1"
names(long_final_cluster)[names(long_final_cluster) ==
"mean_ACT_per2"] <- "Act_Dur_Per2"
names(long_final_cluster)[names(long_final_cluster) ==
"mean_ACT_per3"] <- "Act_Dur_Per3"
names(long_final_cluster)[names(long_final_cluster) ==
"mean_ACT_per4"] <- "Act_Dur_Per4"

Act_Int_Per1 <- long_final_cluster %>% group_by(USUBJID) %>%
   summarise_at(vars(Act_Int_Per1), list(Act_Int_Per1=mean))
Act_Int_Per2 <- long_final_cluster %>% group_by(USUBJID) %>%
   summarise_at(vars(Act_Int_Per2), list(Act_Int_Per2=mean))
Act_Int_Per3 <- long_final_cluster %>% group_by(USUBJID) %>%
   summarise_at(vars(Act_Int_Per3), list(Act_Int_Per3=mean))
Act_Int_Per4 <- long_final_cluster %>% group_by(USUBJID) %>%
   summarise_at(vars(Act_Int_Per4), list(Act_Int_Per4=mean))

Act_Int_Per1$USUBJID <- as.character(Act_Int_Per1$USUBJID)
Act_Int_Per2$USUBJID <- as.character(Act_Int_Per2$USUBJID)
Act_Int_Per3$USUBJID <- as.character(Act_Int_Per3$USUBJID)
Act_Int_Per4$USUBJID <- as.character(Act_Int_Per4$USUBJID)

Act_Dur_Per1 <- long_final_cluster %>% group_by(USUBJID) %>%
   summarise_at(vars(Act_Dur_Per1), list(Act_Dur_Per1=mean))
Act_Dur_Per2 <- long_final_cluster %>% group_by(USUBJID) %>%
   summarise_at(vars(Act_Dur_Per2), list(Act_Dur_Per2=mean))
Act_Dur_Per3 <- long_final_cluster %>% group_by(USUBJID) %>%
   summarise_at(vars(Act_Dur_Per3), list(Act_Dur_Per3=mean))
Act_Dur_Per4 <- long_final_cluster %>% group_by(USUBJID) %>%
   summarise_at(vars(Act_Dur_Per4), list(Act_Dur_Per4=mean))
```

```
Act_Dur_Per1$USUBJID <- as.character(Act_Dur_Per1$USUBJID)
Act_Dur_Per2$USUBJID <- as.character(Act_Dur_Per2$USUBJID)
Act_Dur_Per3$USUBJID <- as.character(Act_Dur_Per3$USUBJID)
Act_Dur_Per4$USUBJID <- as.character(Act_Dur_Per4$USUBJID)

full_4_periods_act_part1 <- merge(Act_Int_Per1,
Act_Int_Per2, by="USUBJID")
full_4_periods_act_part2 <- merge(Act_Int_Per3,
Act_Int_Per4, by="USUBJID")
full_4_periods_act_part3 <- merge(full_4_periods_act_part1,
full_4_periods_act_part2, by="USUBJID")

full_4_periods_act_part4 <- merge(Act_Dur_Per1,
Act_Dur_Per2, by="USUBJID")
full_4_periods_act_part5 <- merge(Act_Dur_Per3,
Act_Dur_Per4, by="USUBJID")
full_4_periods_act_part6 <- merge(full_4_periods_act_part4,
full_4_periods_act_part5, by="USUBJID")

full_4_periods_act_part7 <- merge(full_4_periods_act_part3,
full_4_periods_act_part6, by="USUBJID")
full_4_periods_act_final <- merge(full_4_periods_act_part7,
USUB_cluster, by="USUBJID")

names(full_4_periods_act_final)[names(full_4_periods_act_final)
== "consensuscluster"] <- "Cluster"
full_4_periods_act_final$Cluster =
as.factor(full_4_periods_act_final$Cluster)

full4_table_act_perpat <- table1(~ Act_Int_Per1 + Act_Int_Per2 +
Act_Int_Per3 + Act_Int_Per4 + Act_Dur_Per1 + Act_Dur_Per2
+ Act_Dur_Per3 + Act_Dur_Per4 | Cluster, data=full_4_periods_act_
   final)

full_4_periods_act_final_clin <- merge(full_4_periods_act_final,
advsinmx, by="USUBJID")
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "consensuscluster"] <- "
   Cluster"
full_4_periods_act_final_clin$Cluster
<- as.factor(full_4_periods_act_final_clin$Cluster)
full_4_periods_act_final_clin$CNTYGR1N <- as.character(full_4_
   periods_act_final_clin$CNTYGR1N)
full_4_periods_act_final_clin$CNTYGR1N
[full_4_periods_act_final_clin$CNTYGR1N == "5"]
<- "Eastern␣Europe"
full_4_periods_act_final_clin$CNTYGR1N
[full_4_periods_act_final_clin$CNTYGR1N == "6"]
<- "Western␣Europe␣and␣Israel"
full_4_periods_act_final_clin$CNTYGR1N
[full_4_periods_act_final_clin$CNTYGR1N == "8"]
<- "North␣America"
```

```
full_4_periods_act_final_clin$BASEGFR <- as.character(full_4_
   periods_act_final_clin$BASEGFR)
full_4_periods_act_final_clin$BASEGFR
[full_4_periods_act_final_clin$BASEGFR == ""]
<- "Missing"

full_4_periods_act_final_clin$LVEFHIM <- as.character(full_4_
   periods_act_final_clin$LVEFHIM)
full_4_periods_act_final_clin$LVEFHIM
[full_4_periods_act_final_clin$LVEFHIM == ""]
<- "Missing"

full_4_periods_act_final_clin$AGEGR01 =
as.factor(full_4_periods_act_final_clin$AGEGR01)
full_4_periods_act_final_clin$AGEGR02 =
as.factor(full_4_periods_act_final_clin$AGEGR02)
full_4_periods_act_final_clin$SEX =
as.factor(full_4_periods_act_final_clin$SEX)
full_4_periods_act_final_clin$RACE =
as.factor(full_4_periods_act_final_clin$RACE)
full_4_periods_act_final_clin$CNTYGR1N = as.factor(full_4_periods
   _act_final_clin$CNTYGR1N)
full_4_periods_act_final_clin$NTBNP =
as.factor(full_4_periods_act_final_clin$NTBNP)
full_4_periods_act_final_clin$NYHAFUC =
as.factor(full_4_periods_act_final_clin$NYHAFUC)
full_4_periods_act_final_clin$BETABLK =
as.factor(full_4_periods_act_final_clin$BETABLK)
full_4_periods_act_final_clin$PRIHOSP =
as.factor(full_4_periods_act_final_clin$PRIHOSP)
full_4_periods_act_final_clin$DIABETE =
as.factor(full_4_periods_act_final_clin$DIABETE)

full_4_periods_act_final_clin$ATRFIBR =
as.factor(full_4_periods_act_final_clin$ATRFIBR)
full_4_periods_act_final_clin$HYPERTE =
as.factor(full_4_periods_act_final_clin$HYPERTE)
full_4_periods_act_final_clin$MAXACE =
as.factor(full_4_periods_act_final_clin$MAXACE)
full_4_periods_act_final_clin$MAXARB =
as.factor(full_4_periods_act_final_clin$MAXARB)

full_4_periods_act_final_clin$BASEGFR =
as.factor(full_4_periods_act_final_clin$BASEGFR)
full_4_periods_act_final_clin$BMIGR1 =
as.factor(full_4_periods_act_final_clin$BMIGR1)
full_4_periods_act_final_clin$LVEFB25 =
as.factor(full_4_periods_act_final_clin$LVEFB25)
full_4_periods_act_final_clin$LVEFHIM =
as.factor(full_4_periods_act_final_clin$LVEFHIM)

names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "AGEGR01"]
<- "Age_Group"
```

```
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "CNTYGR1N"]
<- "Country_Group"
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "NYHAFUC"]
<- "NYHA"
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "PRIHOSP"]
<- "Prior_HF_hosp"
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "DIABETE"]
<- "Diabetes"
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "ATRFIBR"]
<- "Afib"
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "HYPERTE"]
<- "Hypertension"
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "BASEGFR"]
<- "EGFR"
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "LVEFHIM"]
<- "LVEF"
names(full_4_periods_act_final_clin)
[names(full_4_periods_act_final_clin) == "BMIGR1"]
<- "BMI"
```

```
df_3_comps <- c("USUBJID", "V2")
df_3_sub <- df_3[df_3_comps]
names(df_3_sub)[names(df_3_sub) == "V2"] <- "Cluster"

baseline_period_clus <-
merge(avg_pat, df_3_sub, by="USUBJID")
baseline_period_act <-
merge(baseline_period_clus, advsinmx, by="USUBJID")

baseline_period_act$Cluster <-
as.factor(baseline_period_act$Cluster)

baseline_period_act$CNTYGR1N <-
as.character(baseline_period_act$CNTYGR1N)

baseline_period_act$CNTYGR1N
[baseline_period_act$CNTYGR1N == "5"] <-
"Eastern␣Europe"
baseline_period_act$CNTYGR1N
[baseline_period_act$CNTYGR1N == "6"] <-
"Western␣Europe␣and␣Israel"
baseline_period_act$CNTYGR1N
[baseline_period_act$CNTYGR1N == "8"] <-
"North␣America"
```

```
baseline_period_act$BASEGFR <- as.character(baseline_period_act$
   BASEGFR)
baseline_period_act$BASEGFR[baseline_period_act$BASEGFR == ""] <-
    "Missing"

baseline_period_act$LVEFHIM <- as.character(baseline_period_act$
   LVEFHIM)
baseline_period_act$LVEFHIM[baseline_period_act$LVEFHIM == ""] <-
    "Missing"

baseline_period_act$AGEGR01 = as.factor(baseline_period_act$
   AGEGR01)
baseline_period_act$AGEGR02 = as.factor(baseline_period_act$
   AGEGR02)
baseline_period_act$SEX = as.factor(baseline_period_act$SEX)
baseline_period_act$RACE = as.factor(baseline_period_act$RACE)
baseline_period_act$CNTYGR1N = as.factor(baseline_period_act$
   CNTYGR1N)
baseline_period_act$NTBNP = as.factor(baseline_period_act$NTBNP)
baseline_period_act$NYHAFUC = as.factor(baseline_period_act$
   NYHAFUC)
baseline_period_act$BETABLK = as.factor(baseline_period_act$
   BETABLK)
baseline_period_act$PRIHOSP = as.factor(baseline_period_act$
   PRIHOSP)
baseline_period_act$DIABETE = as.factor(baseline_period_act$
   DIABETE)
```

```
baseline_period_act$ATRFIBR = as.factor(baseline_period_act$
   ATRFIBR)
baseline_period_act$HYPERTE = as.factor(baseline_period_act$
   HYPERTE)
baseline_period_act$MAXACE = as.factor(baseline_period_act$MAXACE
   )
baseline_period_act$MAXARB = as.factor(baseline_period_act$MAXARB
   )

baseline_period_act$BASEGFR = as.factor(baseline_period_act$
   BASEGFR)
baseline_period_act$BMIGR1 = as.factor(baseline_period_act$BMIGR1
   )
baseline_period_act$LVEFB25 = as.factor(baseline_period_act$
   LVEFB25)
baseline_period_act$LVEFHIM = as.factor(baseline_period_act$
   LVEFHIM)

names(baseline_period_act)[names(baseline_period_act) == "AGEGR01
   "] <- "Age_Group"
names(baseline_period_act)[names(baseline_period_act) == "
   CNTYGR1N"] <- "Country_Group"
names(baseline_period_act)[names(baseline_period_act) == "NYHAFUC
   "] <- "NYHA"
names(baseline_period_act)[names(baseline_period_act) == "PRIHOSP
   "] <- "Prior_HF_hosp"
```

```
names(baseline_period_act)[names(baseline_period_act) == "DIABETE
   "] <- "Diabetes"
names(baseline_period_act)[names(baseline_period_act) == "ATRFIBR
   "] <- "Afib"
names(baseline_period_act)[names(baseline_period_act) == "HYPERTE
   "] <- "Hypertension"
names(baseline_period_act)[names(baseline_period_act) == "BASEGFR
   "] <- "EGFR"
names(baseline_period_act)[names(baseline_period_act) == "LVEFHIM
   "] <- "LVEF"
names(baseline_period_act)[names(baseline_period_act) == "BMIGR1"
   ] <- "BMI"






clin_subset <- subset(adtte, PARAM=="HF␣HOSPITALIZATION" | PARAM
   =="HF␣HOSPITALIZATION␣AND␣URGENT␣VISIT␣FOR␣HF" | PARAM=="
   SECONDARY␣EFFICACY␣OUTCOME", select=c(USUBJID, PARAM, PARAMCD,
    AVAL, CNSR))
table(clin_subset$PARAM, clin_subset$CNSR)
clin_subset_hosp <- subset(adtte, PARAM=="HF␣HOSPITALIZATION" &
   TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣FIRST␣DOSE")
clin_subset_hosp_urg <- subset(adtte, PARAM=="HF␣HOSPITALIZATION␣
   AND␣URGENT␣VISIT␣FOR␣HF" & TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣
   FIRST␣DOSE")
```

```
clin_subset_secondary_eff <- subset(adtte, PARAM=="SECONDARY␣
   EFFICACY␣OUTCOME" & TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣FIRST␣DOSE"
   )
baseline_period_act
myvars <- c("USUBJID", "avg_mean_INT", "avg_mean_ACT", "Cluster",
    "Age_Group", "Country_Group", "NYHA", "Prior_HF_hosp", "
   Diabetes", "Afib", "Hypertension", "EGFR", "BMI")
int_dur_clin_baseline <- baseline_period_act[myvars]

clin_subset_hosp_act <- merge(int_dur_clin_baseline, clin_subset_
   hosp, by="USUBJID")
clin_subset_hosp_act$SEX = as.factor(clin_subset_hosp_act$SEX)
clin_subset_hosp_act$NTBNP = as.factor(clin_subset_hosp_act$NTBNP
   )

clin_subset_hosp_urg_act <- merge(int_dur_clin_baseline, clin_
   subset_hosp_urg, by="USUBJID")
clin_subset_secondary_eff <- merge(int_dur_clin_baseline, clin_
   subset_secondary_eff, by="USUBJID")



Cox Proportional Hazards Model
rm(list = ls())

library(haven)
library(corrplot)
library(caret)
library(epiDisplay)
library(ggplot2)
library(RColorBrewer)
library(plot3D
library(dplyr)
library(reshape2)
library(parallel)
library(xtable)
library(randomForestSRC)
library(ggRandomForests)
library(expss)
library(prodlim)
library(pec)
library(plotly)
library(M3C)
library(table1)
library(expss)
library(survival)
library(survminer)
library(prodlim)
library(pec)
library(plotly)
library(psych)
library(tab)

advsinmx <- read_sas(data_file = ’/adsvinmn_subj.sas7bdat’)
adtte<- read_sas(data_file = ’/adtte.sas7bdat’)
```

```
wide_final<- read_sas(data_file = ’/wide_final.sas7bdat’)
wide_final_baseline <- wide_final[c(23:36)]
df_scale <- scale(wide_final_baseline)
long_final <- read_sas(data_file = ’/final_intraday_part2_ct_dt.
   sas7bdat’)

set.seed(123)
k3 <- kmeans(df_scale, centers = 3, nstart = 25)
str(k3)

USUBJID <- wide_final$USUBJID
id_cluster3 <- cbind(USUBJID, k3$cluster)
final_clusters <- merge(id_cluster3, advsinmx, by="USUBJID")
df_long_3 <- merge(id_cluster3, long_final, by="USUBJID")
final_clusters_step1 <- final_clusters[-c(4:118)]
df_3 <- final_clusters_step1
df_long_3$V2 = as.factor(df_long_3$V2)

describeBy(df_long_3$mean_INT, group=df_long_3$V2, mat=TRUE,
   digits=3)

df_long_3_table <- df_long_3
df_long_3_table_ren <- rename(df_long_3_table, Activity_Intensity
    = mean_INT ,
                       Activity_Duration = mean_ACT)
```

```
df_long_3_table_ren$V3 <- factor(df_long_3_table_ren$V2, levels=c
   (1,2,3), labels=c("Cluster␣1", "Cluster␣2", "Cluster␣3"))

table1(~ Activity_Intensity + Activity_Duration | V2, data=df_
   long_3_table_ren)

df_long_3_table_ren_mrg <- merge(df_long_3_table_ren, advsinmx,
   by="USUBJID")

avg_pat <- read_sas(data_file = '␣/avg_pat_ACT_INT.sas7bdat')
id_cluster3_avg <- merge(id_cluster3, avg_pat)
id_cluster3_avg_merge <- merge(id_cluster3_avg, advsinmx, by="
   USUBJID")
table(df_3$V2)
fviz_cluster(k3, data = df)

df_3$AGEGR01 = as.factor(df_3$AGEGR01)
df_3$AGEGR02 = as.factor(df_3$AGEGR02)
df_3$V2 = as.factor(df_3$V2)

tapply(df_3$AGEGR01, df_3$V2, function(x) table(x)/length(x))
tapply(df_3$AGEGR02, df_3$V2, function(x) table(x)/length(x))

df_long_3$V2 = as.factor(df_long_3$V2)

describeBy(df_long_3$mean_INT, group=df_long_3$V2, mat=TRUE,
   digits=3)
```

```
df_long_3_table <- df_long_3
df_long_3_table_ren <- rename(df_long_3_table, Activity_Intensity
    = mean_INT ,
                       Activity_Duration = mean_ACT)

df_long_3_table_ren$V3 <- factor(df_long_3_table_ren$V2, levels=c
   (1,2,3), labels=c("Cluster␣1", "Cluster␣2", "Cluster␣3"))

table1(~ Activity_Intensity + Activity_Duration | V3, data=df_
   long_3_table_ren)

describeBy(df_long_3$mean_ACT, group=df_long_3$V2, mat=TRUE,
   digits=3)

df_3$SEX = as.factor(df_3$SEX)
table(df_3$V2, df_3$SEX)

tapply(df_3$SEX, df_3$V2, function(x) table(x)/length(x))

df_long_3_table_ren_mrg$SEX = as.factor(df_long_3_table_ren_mrg$
   SEX)
table1(~ Activity_Intensity + Activity_Duration + factor(SEX) |
   V3, data=df_long_3_table_ren_mrg)

df_3$RACE = as.factor(df_3$RACE)
table(df_3$V2, df_3$RACE)
```

```
###Let’s examine CNTYGR1N
df_3$CNTYGR1N = as.factor(df_3$CNTYGR1N)
table(df_3$V2, df_3$CNTYGR1N)

tapply(df_3$CNTYGR1N, df_3$V2, function(x) table(x)/length(x))

df_3$NTBNP = as.factor(df_3$NTBNP)
tapply(df_3$NTBNP, df_3$V2, function(x) table(x)/length(x))

df_3$NYHAFUC = as.factor(df_3$NYHAFUC)
tapply(df_3$NYHAFUC, df_3$V2, function(x) table(x)/length(x))

df_3$BETABLK = as.factor(df_3$BETABLK)
tapply(df_3$BETABLK, df_3$V2, function(x) table(x)/length(x))

df_3$PRIHOSP = as.factor(df_3$PRIHOSP)
tapply(df_3$PRIHOSP, df_3$V2, function(x) table(x)/length(x))

df_3$DIABETE = as.factor(df_3$DIABETE)
tapply(df_3$DIABETE, df_3$V2, function(x) table(x)/length(x))

df_3$ATRFIBR = as.factor(df_3$ATRFIBR)
tapply(df_3$ATRFIBR, df_3$V2, function(x) table(x)/length(x))

df_3$HYPERTE = as.factor(df_3$HYPERTE)
tapply(df_3$HYPERTE, df_3$V2, function(x) table(x)/length(x))
```

```
df_3$MAXACE = as.factor(df_3$MAXACE)
tapply(df_3$MAXACE, df_3$V2, function(x) table(x)/length(x))

df_3$MAXARB = as.factor(df_3$MAXARB)
tapply(df_3$MAXARB, df_3$V2, function(x) table(x)/length(x))

df_3$BASEGFR = as.factor(df_3$BASEGFR)
tapply(df_3$BASEGFR, df_3$V2, function(x) table(x)/length(x))

df_3$BMIGR1 = as.factor(df_3$BMIGR1)
tapply(df_3$BMIGR1, df_3$V2, function(x) table(x)/length(x))

df_3$BMIGR2 = as.factor(df_3$BMIGR2)
tapply(df_3$BMIGR2, df_3$V2, function(x) table(x)/length(x))

df_3$LVEFB25 = as.factor(df_3$LVEFB25)
tapply(df_3$LVEFB25, df_3$V2, function(x) table(x)/length(x))

df_3$LVEFHIM = as.factor(df_3$LVEFHIM)
tapply(df_3$LVEFHIM, df_3$V2, function(x) table(x)/length(x))



df_long_3_table_ren$V3 <- factor(df_long_3_table_ren$V2, levels=c
   (1,2,3), labels=c("Cluster␣1", "Cluster␣2", "Cluster␣3"))
df_long_3_table_ren_mrg$CNTYGR1N <- as.character(df_long_3_table_
   ren_mrg$CNTYGR1N)
```

```
df_long_3_table_ren_mrg$CNTYGR1N[df_long_3_table_ren_mrg$CNTYGR1N
    == "5"] <- "Eastern␣Europe"
df_long_3_table_ren_mrg$CNTYGR1N[df_long_3_table_ren_mrg$CNTYGR1N
    == "6"] <- "Western␣Europe␣and␣Israel"
df_long_3_table_ren_mrg$CNTYGR1N[df_long_3_table_ren_mrg$CNTYGR1N
    == "8"] <- "North␣America"
df_long_3_table_ren_mrg$BASEGFR <- as.character(df_long_3_table_
   ren_mrg$BASEGFR)
df_long_3_table_ren_mrg$LVEFHIM <- as.character(df_long_3_table_
   ren_mrg$LVEFHIM)

df_long_3_table_ren_mrg$AGEGR01 = as.factor(df_long_3_table_ren_
   mrg$AGEGR01)
df_long_3_table_ren_mrg$AGEGR02 = as.factor(df_long_3_table_ren_
   mrg$AGEGR02)
df_long_3_table_ren_mrg$SEX = as.factor(df_long_3_table_ren_mrg$
   SEX)
df_long_3_table_ren_mrg$RACE = as.factor(df_long_3_table_ren_mrg$
   RACE)
df_long_3_table_ren_mrg$CNTYGR1N = as.factor(df_long_3_table_ren_
   mrg$CNTYGR1N)
df_long_3_table_ren_mrg$NTBNP = as.factor(df_long_3_table_ren_mrg
   $NTBNP)
df_long_3_table_ren_mrg$NYHAFUC = as.factor(df_long_3_table_ren_
   mrg$NYHAFUC)
df_long_3_table_ren_mrg$BETABLK = as.factor(df_long_3_table_ren_
   mrg$BETABLK)
```

```
df_long_3_table_ren_mrg$PRIHOSP = as.factor(df_long_3_table_ren_
   mrg$PRIHOSP)
df_long_3_table_ren_mrg$DIABETE = as.factor(df_long_3_table_ren_
   mrg$DIABETE)

df_long_3_table_ren_mrg$ATRFIBR = as.factor(df_long_3_table_ren_
   mrg$ATRFIBR)
df_long_3_table_ren_mrg$HYPERTE = as.factor(df_long_3_table_ren_
   mrg$HYPERTE)
df_long_3_table_ren_mrg$MAXACE = as.factor(df_long_3_table_ren_
   mrg$MAXACE)
df_long_3_table_ren_mrg$MAXARB = as.factor(df_long_3_table_ren_
   mrg$MAXARB)

df_long_3_table_ren_mrg$BASEGFR = as.factor(df_long_3_table_ren_
   mrg$BASEGFR)
df_long_3_table_ren_mrg$BMIGR1 = as.factor(df_long_3_table_ren_
   mrg$BMIGR1)
df_long_3_table_ren_mrg$LVEFB25 = as.factor(df_long_3_table_ren_
   mrg$LVEFB25)
df_long_3_table_ren_mrg$LVEFHIM = as.factor(df_long_3_table_ren_
   mrg$LVEFHIM)
table1(~ Activity_Intensity + Activity_Duration + factor(AGEGR01)
    + factor(AGEGR02) + factor(SEX) + factor(CNTYGR1N) + factor(
   NTBNP) + factor(NYHAFUC) + factor(PRIHOSP) + factor(DIABETE) +
    factor(ATRFIBR)
```

```
     + factor(HYPERTE) + factor(BASEGFR) + factor(BMIGR1) +
        factor(LVEFB25) + factor(LVEFHIM) | V2, data=df_long_3_
        table_ren_mrg)
id_cluster3_avg_merge$AGEGR01 = as.factor(id_cluster3_avg_merge$
   AGEGR01)
id_cluster3_avg_merge$AGEGR02 = as.factor(id_cluster3_avg_merge$
   AGEGR02)
id_cluster3_avg_merge$SEX = as.factor(id_cluster3_avg_merge$SEX)
id_cluster3_avg_merge$RACE = as.factor(id_cluster3_avg_merge$RACE
   )
id_cluster3_avg_merge$CNTYGR1N = as.factor(id_cluster3_avg_merge$
   CNTYGR1N)
id_cluster3_avg_merge$NTBNP = as.factor(id_cluster3_avg_merge$
   NTBNP)
id_cluster3_avg_merge$NYHAFUC = as.factor(id_cluster3_avg_merge$
   NYHAFUC)
id_cluster3_avg_merge$BETABLK = as.factor(id_cluster3_avg_merge$
   BETABLK)
id_cluster3_avg_merge$PRIHOSP = as.factor(id_cluster3_avg_merge$
   PRIHOSP)
id_cluster3_avg_merge$DIABETE = as.factor(id_cluster3_avg_merge$
   DIABETE)

id_cluster3_avg_merge$ATRFIBR = as.factor(id_cluster3_avg_merge$
   ATRFIBR)
id_cluster3_avg_merge$HYPERTE = as.factor(id_cluster3_avg_merge$
   HYPERTE)
```

```
id_cluster3_avg_merge$MAXACE = as.factor(id_cluster3_avg_merge$
   MAXACE)
id_cluster3_avg_merge$MAXARB = as.factor(id_cluster3_avg_merge$
   MAXARB)

id_cluster3_avg_merge$BASEGFR = as.factor(id_cluster3_avg_merge$
   BASEGFR)
id_cluster3_avg_merge$BMIGR1 = as.factor(id_cluster3_avg_merge$
   BMIGR1)
id_cluster3_avg_merge$LVEFB25 = as.factor(id_cluster3_avg_merge$
   LVEFB25)
id_cluster3_avg_merge$LVEFHIM = as.factor(id_cluster3_avg_merge$
   LVEFHIM)

names(id_cluster3_avg_merge)[names(id_cluster3_avg_merge) == "avg
   _mean_INT"] <- "Activity_Intensity"
names(id_cluster3_avg_merge)[names(id_cluster3_avg_merge) == "avg
   _mean_ACT"] <- "Activity_Duration"

table1(~ Activity_Intensity + Activity_Duration + factor(AGEGR01)
    + factor(AGEGR02) + factor(SEX) + factor(CNTYGR1N) + factor(
   NTBNP) + factor(NYHAFUC) + factor(PRIHOSP) + factor(DIABETE) +
    factor(ATRFIBR)
     + factor(HYPERTE) + factor(BASEGFR) + factor(BMIGR1) +
        factor(LVEFB25) + factor(LVEFHIM) | V2, data=id_cluster3_
        avg_merge)
id_cluster3_avg <- merge(id_cluster3, avg_pat)
```

```
id_cluster3_avg_merge <- merge(id_cluster3_avg, advsinmx, by="
   USUBJID")
id_cluster3_avg_merge <- id_cluster3_avg_merge[!(id_cluster3_avg_
   merge$NTBNP=="Missing"),]
id_cluster3_avg_merge <- id_cluster3_avg_merge[!(id_cluster3_avg_
   merge$BASEGFR==""),]

id_cluster3_ren <- rename(id_cluster3_avg_merge, Activity_
   Intensity = avg_mean_INT, Activity_Duration = avg_mean_ACT)
id_cluster3_ren$V3 <- factor(id_cluster3_ren$V2, levels=c(1,2,3),
    labels=c("Cluster␣1", "Cluster␣2", "Cluster␣3"))
id_cluster3_ren_mrg <- id_cluster3_ren
id_cluster3_ren_mrg$CNTYGR1N <- as.character(id_cluster3_ren_mrg$
   CNTYGR1N)
id_cluster3_ren_mrg$CNTYGR1N[id_cluster3_ren_mrg $CNTYGR1N == "5"
   ] <- "Eastern␣Europe"
id_cluster3_ren_mrg$CNTYGR1N[id_cluster3_ren_mrg $CNTYGR1N == "6"
   ] <- "Western␣Europe␣and␣Israel"
id_cluster3_ren_mrg$CNTYGR1N[id_cluster3_ren_mrg $CNTYGR1N == "8"
   ] <- "North␣America"
id_cluster3_ren_mrg$BASEGFR <- as.character(id_cluster3_ren_mrg $
   BASEGFR)
id_cluster3_ren_mrg$LVEFHIM <- as.character(id_cluster3_ren_mrg $
   LVEFHIM)
id_cluster3_ren_mrg$AGEGR01 = as.factor(id_cluster3_ren_mrg$
   AGEGR01)
```

```
id_cluster3_ren_mrg$AGEGR02 = as.factor(id_cluster3_ren_mrg$
   AGEGR02)
id_cluster3_ren_mrg$SEX = as.factor(id_cluster3_ren_mrg$SEX)
id_cluster3_ren_mrg$RACE = as.factor(id_cluster3_ren_mrg$RACE)
id_cluster3_ren_mrg$CNTYGR1N = as.factor(id_cluster3_ren_mrg$
   CNTYGR1N)
id_cluster3_ren_mrg$NTBNP = as.factor(id_cluster3_ren_mrg$NTBNP)
id_cluster3_ren_mrg$NYHAFUC = as.factor(id_cluster3_ren_mrg$
   NYHAFUC)
id_cluster3_ren_mrg$BETABLK = as.factor(id_cluster3_ren_mrg$
   BETABLK)
id_cluster3_ren_mrg$PRIHOSP = as.factor(id_cluster3_ren_mrg$
   PRIHOSP)
id_cluster3_ren_mrg$DIABETE = as.factor(id_cluster3_ren_mrg$
   DIABETE)
id_cluster3_ren_mrg$ATRFIBR = as.factor(id_cluster3_ren_mrg$
   ATRFIBR)
id_cluster3_ren_mrg$HYPERTE = as.factor(id_cluster3_ren_mrg$
   HYPERTE)
id_cluster3_ren_mrg$MAXACE = as.factor(id_cluster3_ren_mrg$MAXACE
   )
id_cluster3_ren_mrg$MAXARB = as.factor(id_cluster3_ren_mrg$MAXARB
   )
id_cluster3_ren_mrg$BASEGFR = as.factor(id_cluster3_ren_mrg$
   BASEGFR)
id_cluster3_ren_mrg$BMIGR1 = as.factor(id_cluster3_ren_mrg$BMIGR1
   )
```

```
id_cluster3_ren_mrg$LVEFB25 = as.factor(id_cluster3_ren_mrg$
   LVEFB25)a
id_cluster3_ren_mrg$LVEFHIM = as.factor(id_cluster3_ren_mrg$
   LVEFHIM)

names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "AGEGR01
   "] <- "Age_Group"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "
   CNTYGR1N"] <- "Country_Group"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "NYHAFUC
   "] <- "NYHA"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "PRIHOSP
   "] <- "Prior_HF_hosp"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "DIABETE
   "] <- "Diabetes"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "ATRFIBR
   "] <- "Afib"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "HYPERTE
   "] <- "Hypertension"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "BASEGFR
   "] <- "EGFR"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "LVEFHIM
   "] <- "LVEF"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "BMIGR1"
   ] <- "BMI"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "V2"] <-
    "Cluster"
```

```
final_table <- table1(~ Activity_Intensity + Activity_Duration +
   factor(Age_Group) + factor(SEX) + factor(Country_Group) +
   factor(NTBNP) + factor(NYHA) + factor(Prior_HF_hosp) + factor(
   Diabetes) + factor(Afib) + factor(Hypertension) + factor(EGFR)
    + factor(BMI) | Cluster, data=id_cluster3_ren_mrg)
main_RSF_vars <- id_cluster3_ren_mrg[var_interest]
main_RSF_vars <- main_RSF_vars[, -16]
clin_subset <- subset(adtte, PARAM=="HF␣HOSPITALIZATION" | PARAM
   =="HF␣HOSPITALIZATION␣AND␣URGENT␣VISIT␣FOR␣HF" | PARAM=="
   SECONDARY␣EFFICACY␣OUTCOME", select=c(USUBJID, PARAM, PARAMCD,
    AVAL, CNSR))
table(clin_subset$PARAM, clin_subset$CNSR)
clin_subset_hosp <- subset(adtte, PARAM=="HF␣HOSPITALIZATION" &
   TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣FIRST␣DOSE")
clin_subset_hosp_urg <- subset(adtte, PARAM=="HF␣HOSPITALIZATION␣
   AND␣URGENT␣VISIT␣FOR␣HF" & TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣
   FIRST␣DOSE")
clin_subset_secondary_eff <- subset(adtte, PARAM=="SECONDARY␣
   EFFICACY␣OUTCOME" & TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣FIRST␣DOSE"
   )
hosp_cnsr <- c("USUBJID", "AVAL", "CNSR")
clin_subset_hosp_cnsr <- clin_subset_hosp[hosp_cnsr]

RSF_hosp_final <- merge(main_RSF_vars, clin_subset_hosp_cnsr, by=
   "USUBJID")
RSF_hosp_final <- RSF_hosp_final[,-1]
RSF_hosp_final <- RSF_hosp_final[,-3]
```

```
RSF_hosp_final$CNSR2 <- ifelse(RSF_hosp_final$CNSR==0, 1, 0)
RSF_hosp_final <- RSF_hosp_final[, -c(15)]
names(RSF_hosp_final)
coxPH_hosp_final <- RSF_hosp_final

res.cox.hosp_BWS <- selectCox(Surv(AVAL, CNSR2) ~ Activity_
   Intensity + Activity_Duration + Age_Group + SEX + Country_
   Group + NYHA + NTBNP + Prior_HF_hosp + Diabetes + Afib +
   Hypertension + EGFR + BMI, data=coxPH_hosp_final)
res.cox.hosp <- coxph(Surv(AVAL, CNSR2) ~ Activity_Intensity +
   Activity_Duration + Age_Group + SEX + Country_Group + NYHA +
   NTBNP + Prior_HF_hosp + Diabetes + Afib + Hypertension + EGFR
   + BMI, data=coxPH_hosp_final)
res.cox.hosp_pub <- coxphSeries(Surv(AVAL, CNSR2==1) ~ Activity_
   Intensity + Activity_Duration + Age_Group + SEX + Country_
   Group + NYHA + NTBNP + Prior_HF_hosp + Diabetes + Afib +
   Hypertension + EGFR + BMI, vars=c("Activity_Intensity", "
   Activity_Duration", "Age_Group", "SEX", "Country_Group", "NYHA
   ", "NTBNP", "Prior_HF_hosp", "Diabetes", "Afib", "Hypertension
   ", "EGFR", "BMI"), data=coxPH_hosp_final)
publish(res.cox.hosp_pub)
dev.off(dev.list()["RStudioGD"])
grid.hosp <- grid.table(res.cox.hosp_pub)
res.cox.hosp_final <- coxph(Surv(AVAL, CNSR2) ~NTBNP+NYHA, data=
   coxPH_hosp_final)
res.cox.hosp_pub_final <- coxphSeries(Surv(AVAL, CNSR2==1) ~ NYHA
    + NTBNP, vars=c("NYHA", "NTBNP"), data=coxPH_hosp_final)
```

```
publish(res.cox.hosp_pub_final)
dev.off(dev.list()["RStudioGD"])
grid.hosp <- grid.table(res.cox.hosp_pub_final)

test.ph = cox.zph(res.cox.hosp_final)
test.ph
publish(test.ph)
grid.hosp <- grid.table(test.ph)
ggcoxzph(test.ph)

hosp_urg_cnsr <- c("USUBJID", "AVAL", "CNSR")
clin_subset_hosp_urg_cnsr <- clin_subset_hosp_urg[hosp_urg_cnsr]
RSF_hosp_urg_final <- merge(main_RSF_vars, clin_subset_hosp_urg_
   cnsr, by="USUBJID")
RSF_hosp_urg_final <- RSF_hosp_urg_final[,-1]
RSF_hosp_urg_final <- RSF_hosp_urg_final[,-3]
RSF_hosp_urg_final$CNSR2 <- ifelse(RSF_hosp_urg_final$CNSR==0, 1,
    0)
RSF_hosp_urg_final <- RSF_hosp_urg_final[, -c(16)]
RSF_hosp_urg_final <- RSF_hosp_urg_final[, -c(15)]
names(RSF_hosp_urg_final)

coxPH_hosp_urg_final <- RSF_hosp_urg_final
res.cox.hosp_urg_SEX <- coxph(Surv(AVAL, CNSR2) ~ SEX, data=coxPH
   _hosp_urg_final) #nonsignificant
res.cox.hosp_urg_CG <- coxph(Surv(AVAL, CNSR2) ~ Country_Group,
   data=coxPH_hosp_urg_final) #nonsignificant
```

```
res.cox.hosp_urg_NYHA <- coxph(Surv(AVAL, CNSR2) ~ NYHA, data=
   coxPH_hosp_urg_final) #significant
res.cox.hosp_urg_NT <- coxph(Surv(AVAL, CNSR2) ~ NTBNP, data=
   coxPH_hosp_urg_final) #significant
res.cox.hosp_urg_PH <- coxph(Surv(AVAL, CNSR2) ~ Prior_HF_hosp,
   data=coxPH_hosp_urg_final) #nonsignificant
res.cox.hosp_urg_DB <- coxph(Surv(AVAL, CNSR2) ~ Diabetes, data=
   coxPH_hosp_urg_final) #barely non significant
res.cox.hosp_urg_AF <- coxph(Surv(AVAL, CNSR2) ~ Afib, data=coxPH
   _hosp_urg_final) #significant
res.cox.hosp_urg_HT <- coxph(Surv(AVAL, CNSR2) ~ Hypertension,
   data=coxPH_hosp_urg_final) #nonsignificant
res.cox.hosp_urg_EG <- coxph(Surv(AVAL, CNSR2) ~ EGFR, data=coxPH
   _hosp_urg_final) #significant
res.cox.hosp_urg_BM <- coxph(Surv(AVAL, CNSR2) ~ BMI, data=coxPH_
   hosp_urg_final) #nonsignificant
res.cox.hosp_urg_pub <- coxphSeries(Surv(AVAL, CNSR2==1) ~
   Activity_Intensity + Activity_Duration + Age_Group + SEX +
   Country_Group + NYHA + NTBNP + Prior_HF_hosp + Diabetes + Afib
    + Hypertension + EGFR + BMI, vars=c("Activity_Intensity", "
   Activity_Duration", "Age_Group", "SEX", "Country_Group", "NYHA
   ", "NTBNP", "Prior_HF_hosp", "Diabetes", "Afib", "Hypertension
   ", "EGFR", "BMI"), data=coxPH_hosp_urg_final)
publish(res.cox.hosp_urg_pub)
dev.off(dev.list()["RStudioGD"])
grid.hosp_urg <- grid.table(res.cox.hosp_urg_pub)
```

```
res.cox.hosp_final <- coxph(Surv(AVAL, CNSR2) ~NTBNP+NYHA, data=
   coxPH_hosp_final)
res.cox.hosp_urg_BWS <- selectCox(Surv(AVAL, CNSR2) ~ Activity_
   Intensity + Activity_Duration + Age_Group + SEX + Country_
   Group + NYHA + NTBNP + Prior_HF_hosp + Diabetes + Afib +
   Hypertension + EGFR + BMI, data=coxPH_hosp_urg_final)
res.cox.urg.hosp <- coxph(Surv(AVAL, CNSR2) ~ Activity_Intensity
   + Activity_Duration + Age_Group + SEX + Country_Group + NYHA +
    NTBNP + Prior_HF_hosp + Diabetes + Afib + Hypertension + EGFR
    + BMI, data=coxPH_hosp_urg_final)
res.cox.urg.hosp_final <- coxph(Surv(AVAL, CNSR2) ~NTBNP+NYHA,
   data=coxPH_hosp_urg_final)
res.cox.hosp_urg_pub_final <- coxphSeries(Surv(AVAL, CNSR2==1) ~
   NYHA + NTBNP, vars=c("NYHA", "NTBNP"), data=coxPH_hosp_urg_
   final)
publish(res.cox.hosp_urg_pub_final)
dev.off(dev.list()["RStudioGD"])
grid.hosp <- grid.table(res.cox.hosp_urg_pub_final)
test.ph = cox.zph(res.cox.urg.hosp_final)
test.ph
ggcoxzph(test.ph)

clin_subset_secondary_eff <- subset(adtte, PARAM=="SECONDARY␣
   EFFICACY␣OUTCOME" & TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣FIRST␣DOSE"
   )
secondary_cnsr <- c("USUBJID", "AVAL", "CNSR")
clin_subset_sec_cnsr <- clin_subset_secondary_eff[secondary_cnsr]
```

```
RSF_sec_final <- merge(main_RSF_vars, clin_subset_sec_cnsr, by="
   USUBJID")
RSF_sec_final <- RSF_sec_final[,-c(1, 4, 18)]
RSF_sec_final$CNSR2 <- ifelse(RSF_sec_final$CNSR==0, 1, 0)
table(RSF_sec_final$CNSR)
RSF_sec_final <- RSF_sec_final[, -c(15)]
names(RSF_sec_final)

coxPH_sec_final <- RSF_sec_final
res.cox.sec_AI <- coxph(Surv(AVAL, CNSR2) ~ Activity_Intensity,
   data=coxPH_sec_final) #significant
res.cox.sec_AD <- coxph(Surv(AVAL, CNSR2) ~ Activity_Duration,
   data=coxPH_sec_final) #significant
res.cox.sec_SEX <- coxph(Surv(AVAL, CNSR2) ~ SEX, data=coxPH_sec_
   final) #nonsignificant
res.cox.sec_CG <- coxph(Surv(AVAL, CNSR2) ~ Country_Group, data=
   coxPH_sec_final) #nonsignificant
res.cox.sec_NYHA <- coxph(Surv(AVAL, CNSR2) ~ NYHA, data=coxPH_
   sec_final) #significant
res.cox.sec_NT <- coxph(Surv(AVAL, CNSR2) ~ NTBNP, data=coxPH_sec
   _final) #significant
res.cox.sec_PH <- coxph(Surv(AVAL, CNSR2) ~ Prior_HF_hosp, data=
   coxPH_sec_final) #barely nonsignificant
res.cox.sec_DB <- coxph(Surv(AVAL, CNSR2) ~ Diabetes, data=coxPH_
   sec_final) #nonsignificant
```

```
res.cox.sec_AF <- coxph(Surv(AVAL, CNSR2) ~ Afib, data=coxPH_sec_
   final) #significant
res.cox.sec_HT <- coxph(Surv(AVAL, CNSR2) ~ Hypertension, data=
   coxPH_sec_final) #nonsignificant
res.cox.sec_EG <- coxph(Surv(AVAL, CNSR2) ~ EGFR, data=coxPH_sec_
   final) #significant
res.cox.sec_BM <- coxph(Surv(AVAL, CNSR2) ~ BMI, data=coxPH_sec_
   final) #nonsignificant

res.cox.sec_BWS <- selectCox(Surv(AVAL, CNSR2) ~ Activity_
   Intensity + Activity_Duration + Age_Group + SEX + Country_
   Group + NYHA + NTBNP + Prior_HF_hosp + Diabetes + Afib +
   Hypertension + EGFR + BMI, data=coxPH_sec_final)

res.cox.sec_pub <- coxphSeries(Surv(AVAL, CNSR2==1) ~ Activity_
   Intensity + Activity_Duration + Age_Group + SEX + Country_
   Group + NYHA + NTBNP + Prior_HF_hosp + Diabetes + Afib +
   Hypertension + EGFR + BMI, vars=c("Activity_Intensity", "
   Activity_Duration", "Age_Group", "SEX", "Country_Group", "NYHA
   ", "NTBNP", "Prior_HF_hosp", "Diabetes", "Afib", "Hypertension
   ", "EGFR", "BMI"), data=coxPH_sec_final)
publish(res.cox.sec_pub)
dev.off(dev.list()["RStudioGD"])
grid.hosp_urg <- grid.table(res.cox.sec_pub)
res.cox.sec <- coxph(Surv(AVAL, CNSR2) ~ Activity_Intensity +
   Activity_Duration + Age_Group + SEX + Country_Group + NYHA +
   NTBNP + Prior_HF_hosp + Diabetes + Afib + Hypertension + EGFR
   + BMI, data=coxPH_sec_final)
res.cox.sec_final <- coxph(Surv(AVAL, CNSR2) ~NTBNP+NYHA, data=
   coxPH_sec_final)
res.cox.sec_pub_final <- coxphSeries(Surv(AVAL, CNSR2==1) ~ NYHA
   + NTBNP, vars=c("NYHA", "NTBNP"), data=coxPH_sec_final)
publish(res.cox.sec_pub_final)
dev.off(dev.list()["RStudioGD"])
grid.hosp <- grid.table(res.cox.sec_pub_final)

test.ph = cox.zph(res.cox.sec_final)
ggcoxzph(test.ph)







Random Survival Forest
rm(list = ls())

library(haven)
library(corrplot)
library(caret)
library(epiDisplay)
library(ggplot2)
library(RColorBrewer)
library(plot3D)
library(dplyr)
library(reshape2)
library(parallel)
library(xtable)
library(randomForestSRC)
library(ggRandomForests)
library(expss)
library(survival)
library(prodlim)
library(pec)
library(plotly)
library(M3C)
library(table1)
library(haven)
library(corrplot)
library(caret)
library(epiDisplay)
advsinmx <- read_sas(data_file = ’/adsvinmn_subj.sas7bdat’)
adtte<- read_sas(data_file = ’/adtte.sas7bdat’)
wide_final<- read_sas(data_file = ’/wide_final.sas7bdat’)
long_final <- read_sas(data_file = ’/final_intraday_part2_ct_dt.
   sas7bdat’)
wide_final_baseline <- wide_final[c(23:36)]
df_scale <- scale(wide_final_baseline)
set.seed(123)
k3 <- kmeans(df_scale, centers = 3, nstart = 25)
str(k3)
```

```
USUBJID <- wide_final$USUBJID
id_cluster3 <- cbind(USUBJID, k3$cluster)
final_clusters <- merge(id_cluster3, advsinmx, by="USUBJID")

df_long_3 <- merge(id_cluster3, long_final, by="USUBJID")
final_clusters_step1 <- final_clusters[-c(4:118)]

df_3 <- final_clusters_step1
df_long_3$V2 = as.factor(df_long_3$V2)
describeBy(df_long_3$mean_INT, group=df_long_3$V2, mat=TRUE,
   digits=3)

df_long_3_table <- df_long_3
df_long_3_table_ren <- rename(df_long_3_table, Activity_Intensity
    = mean_INT ,
                      Activity_Duration = mean_ACT)

df_long_3_table_ren$V3 <- factor(df_long_3_table_ren$V2, levels=c
   (1,2,3), labels=c("Cluster␣1", "Cluster␣2", "Cluster␣3"))

table1(~ Activity_Intensity + Activity_Duration | V2, data=df_
   long_3_table_ren)
df_long_3_table_ren_mrg <- merge(df_long_3_table_ren, advsinmx,
   by="USUBJID")
avg_pat <- read_sas(data_file = '/avg_pat_ACT_INT.sas7bdat')
id_cluster3_avg <- merge(id_cluster3, avg_pat)
```

```
id_cluster3_avg_merge <- merge(id_cluster3_avg, advsinmx, by="
   USUBJID")

table(df_3$V2)
df_3$AGEGR01 = as.factor(df_3$AGEGR01)
df_3$AGEGR02 = as.factor(df_3$AGEGR02)
df_3$V2 = as.factor(df_3$V2)

tapply(df_3$AGEGR01, df_3$V2, function(x) table(x)/length(x))
tapply(df_3$AGEGR02, df_3$V2, function(x) table(x)/length(x))
df_long_3$V2 = as.factor(df_long_3$V2)

describeBy(df_long_3$mean_INT, group=df_long_3$V2, mat=TRUE,
   digits=3)

df_long_3_table <- df_long_3
df_long_3_table_ren <- rename(df_long_3_table, Activity_Intensity
    = mean_INT ,
                              Activity_Duration = mean_ACT)

df_long_3_table_ren$V3 <- factor(df_long_3_table_ren$V2, levels=c
   (1,2,3), labels=c("Cluster␣1", "Cluster␣2", "Cluster␣3"))

table1(~ Activity_Intensity + Activity_Duration | V3, data=df_
   long_3_table_ren)
describeBy(df_long_3$mean_ACT, group=df_long_3$V2, mat=TRUE,
   digits=3)
```

```
df_3$SEX = as.factor(df_3$SEX)
table(df_3$V2, df_3$SEX)

tapply(df_3$SEX, df_3$V2, function(x) table(x)/length(x))

df_long_3_table_ren_mrg$SEX = as.factor(df_long_3_table_ren_mrg$
   SEX)
table1(~ Activity_Intensity + Activity_Duration + factor(SEX) |
   V3, data=df_long_3_table_ren_mrg)

df_3$RACE = as.factor(df_3$RACE)
table(df_3$V2, df_3$RACE)
df_3$CNTYGR1N = as.factor(df_3$CNTYGR1N)
table(df_3$V2, df_3$CNTYGR1N)

tapply(df_3$CNTYGR1N, df_3$V2, function(x) table(x)/length(x))
df_3$NTBNP = as.factor(df_3$NTBNP)
tapply(df_3$NTBNP, df_3$V2, function(x) table(x)/length(x))
df_3$NYHAFUC = as.factor(df_3$NYHAFUC)
tapply(df_3$NYHAFUC, df_3$V2, function(x) table(x)/length(x))

df_3$BETABLK = as.factor(df_3$BETABLK)
tapply(df_3$BETABLK, df_3$V2, function(x) table(x)/length(x))
df_3$PRIHOSP = as.factor(df_3$PRIHOSP)
tapply(df_3$PRIHOSP, df_3$V2, function(x) table(x)/length(x))
df_3$DIABETE = as.factor(df_3$DIABETE)
tapply(df_3$DIABETE, df_3$V2, function(x) table(x)/length(x))
```

```
df_3$ATRFIBR = as.factor(df_3$ATRFIBR)
tapply(df_3$ATRFIBR, df_3$V2, function(x) table(x)/length(x))
df_3$HYPERTE = as.factor(df_3$HYPERTE)
tapply(df_3$HYPERTE, df_3$V2, function(x) table(x)/length(x))
df_3$MAXACE = as.factor(df_3$MAXACE)
tapply(df_3$MAXACE, df_3$V2, function(x) table(x)/length(x))

df_3$MAXARB = as.factor(df_3$MAXARB)
tapply(df_3$MAXARB, df_3$V2, function(x) table(x)/length(x))

df_3$BASEGFR = as.factor(df_3$BASEGFR)
tapply(df_3$BASEGFR, df_3$V2, function(x) table(x)/length(x))

df_3$BMIGR1 = as.factor(df_3$BMIGR1)
tapply(df_3$BMIGR1, df_3$V2, function(x) table(x)/length(x))
df_3$BMIGR2 = as.factor(df_3$BMIGR2)
tapply(df_3$BMIGR2, df_3$V2, function(x) table(x)/length(x))
df_3$LVEFB25 = as.factor(df_3$LVEFB25)
tapply(df_3$LVEFB25, df_3$V2, function(x) table(x)/length(x))
df_3$LVEFHIM = as.factor(df_3$LVEFHIM)
tapply(df_3$LVEFHIM, df_3$V2, function(x) table(x)/length(x))

df_long_3_table_ren$V3 <- factor(df_long_3_table_ren$V2, levels=c
   (1,2,3), labels=c("Cluster␣1", "Cluster␣2", "Cluster␣3"))
df_long_3_table_ren_mrg$CNTYGR1N <- as.character(df_long_3_table_
   ren_mrg$CNTYGR1N)
```

```
df_long_3_table_ren_mrg$CNTYGR1N[df_long_3_table_ren_mrg$CNTYGR1N
    == "5"] <- "Eastern␣Europe"
df_long_3_table_ren_mrg$CNTYGR1N[df_long_3_table_ren_mrg$CNTYGR1N
    == "6"] <- "Western␣Europe␣and␣Israel"
df_long_3_table_ren_mrg$CNTYGR1N[df_long_3_table_ren_mrg$CNTYGR1N
    == "8"] <- "North␣America"
df_long_3_table_ren_mrg$BASEGFR <- as.character(df_long_3_table_
   ren_mrg$BASEGFR)
df_long_3_table_ren_mrg$LVEFHIM <- as.character(df_long_3_table_
   ren_mrg$LVEFHIM)
df_long_3_table_ren_mrg$AGEGR01 = as.factor(df_long_3_table_ren_
   mrg$AGEGR01)
df_long_3_table_ren_mrg$AGEGR02 = as.factor(df_long_3_table_ren_
   mrg$AGEGR02)
df_long_3_table_ren_mrg$SEX = as.factor(df_long_3_table_ren_mrg$
   SEX)
df_long_3_table_ren_mrg$RACE = as.factor(df_long_3_table_ren_mrg$
   RACE)
df_long_3_table_ren_mrg$CNTYGR1N = as.factor(df_long_3_table_ren_
   mrg$CNTYGR1N)
df_long_3_table_ren_mrg$NTBNP = as.factor(df_long_3_table_ren_mrg
   $NTBNP)
df_long_3_table_ren_mrg$NYHAFUC = as.factor(df_long_3_table_ren_
   mrg$NYHAFUC)
df_long_3_table_ren_mrg$BETABLK = as.factor(df_long_3_table_ren_
   mrg$BETABLK)
```

```
df_long_3_table_ren_mrg$PRIHOSP = as.factor(df_long_3_table_ren_
   mrg$PRIHOSP)
df_long_3_table_ren_mrg$DIABETE = as.factor(df_long_3_table_ren_
   mrg$DIABETE)
df_long_3_table_ren_mrg$ATRFIBR = as.factor(df_long_3_table_ren_
   mrg$ATRFIBR)
df_long_3_table_ren_mrg$HYPERTE = as.factor(df_long_3_table_ren_
   mrg$HYPERTE)
df_long_3_table_ren_mrg$MAXACE = as.factor(df_long_3_table_ren_
   mrg$MAXACE)
df_long_3_table_ren_mrg$MAXARB = as.factor(df_long_3_table_ren_
   mrg$MAXARB)
df_long_3_table_ren_mrg$BASEGFR = as.factor(df_long_3_table_ren_
   mrg$BASEGFR)
df_long_3_table_ren_mrg$BMIGR1 = as.factor(df_long_3_table_ren_
   mrg$BMIGR1)
df_long_3_table_ren_mrg$LVEFB25 = as.factor(df_long_3_table_ren_
   mrg$LVEFB25)
df_long_3_table_ren_mrg$LVEFHIM = as.factor(df_long_3_table_ren_
   mrg$LVEFHIM)
table1(~ Activity_Intensity + Activity_Duration + factor(AGEGR01)
    + factor(AGEGR02) + factor(SEX) + factor(CNTYGR1N) + factor(
   NTBNP) + factor(NYHAFUC) + factor(PRIHOSP) + factor(DIABETE) +
    factor(ATRFIBR) + factor(HYPERTE) + factor(BASEGFR) + factor(
   BMIGR1) + factor(LVEFB25) + factor(LVEFHIM) | V2, data=df_long
   _3_table_ren_mrg)
```

```
id_cluster3_avg_merge$AGEGR01 = as.factor(id_cluster3_avg_merge$
   AGEGR01)
id_cluster3_avg_merge$AGEGR02 = as.factor(id_cluster3_avg_merge$
   AGEGR02)
id_cluster3_avg_merge$SEX = as.factor(id_cluster3_avg_merge$SEX)
id_cluster3_avg_merge$RACE = as.factor(id_cluster3_avg_merge$RACE
   )
id_cluster3_avg_merge$CNTYGR1N = as.factor(id_cluster3_avg_merge$
   CNTYGR1N)
id_cluster3_avg_merge$NTBNP = as.factor(id_cluster3_avg_merge$
   NTBNP)
id_cluster3_avg_merge$NYHAFUC = as.factor(id_cluster3_avg_merge$
   NYHAFUC)
id_cluster3_avg_merge$BETABLK = as.factor(id_cluster3_avg_merge$
   BETABLK)
id_cluster3_avg_merge$PRIHOSP = as.factor(id_cluster3_avg_merge$
   PRIHOSP)
id_cluster3_avg_merge$DIABETE = as.factor(id_cluster3_avg_merge$
   DIABETE)
id_cluster3_avg_merge$ATRFIBR = as.factor(id_cluster3_avg_merge$
   ATRFIBR)
id_cluster3_avg_merge$HYPERTE = as.factor(id_cluster3_avg_merge$
   HYPERTE)
id_cluster3_avg_merge$MAXACE = as.factor(id_cluster3_avg_merge$
   MAXACE)
id_cluster3_avg_merge$MAXARB = as.factor(id_cluster3_avg_merge$
   MAXARB)
```

```
id_cluster3_avg_merge$BASEGFR = as.factor(id_cluster3_avg_merge$
   BASEGFR)
id_cluster3_avg_merge$BMIGR1 = as.factor(id_cluster3_avg_merge$
   BMIGR1)
id_cluster3_avg_merge$LVEFB25 = as.factor(id_cluster3_avg_merge$
   LVEFB25)
id_cluster3_avg_merge$LVEFHIM = as.factor(id_cluster3_avg_merge$
   LVEFHIM)

names(id_cluster3_avg_merge)[names(id_cluster3_avg_merge) == "avg
   _mean_INT"] <- "Activity_Intensity"
names(id_cluster3_avg_merge)[names(id_cluster3_avg_merge) == "avg
   _mean_ACT"] <- "Activity_Duration"

table1(~ Activity_Intensity + Activity_Duration + factor(AGEGR01)
    + factor(AGEGR02) + factor(SEX) + factor(CNTYGR1N) + factor(
   NTBNP) + factor(NYHAFUC) + factor(PRIHOSP) + factor(DIABETE) +
    factor(ATRFIBR) + factor(HYPERTE) + factor(BASEGFR) + factor(
   BMIGR1) + factor(LVEFB25) + factor(LVEFHIM) | V2, data=id_
   cluster3_avg_merge)

id_cluster3_avg <- merge(id_cluster3, avg_pat)
id_cluster3_avg_merge <- merge(id_cluster3_avg, advsinmx, by="
   USUBJID")
id_cluster3_ren <- rename(id_cluster3_avg_merge, Activity_
   Intensity = avg_mean_INT, Activity_Duration = avg_mean_ACT)
```

```
id_cluster3_ren$V3 <- factor(id_cluster3_ren$V2, levels=c(1,2,3),
    labels=c("Cluster␣1", "Cluster␣2", "Cluster␣3"))
id_cluster3_ren_mrg <- id_cluster3_ren
id_cluster3_ren_mrg$CNTYGR1N <- as.character(id_cluster3_ren_mrg$
   CNTYGR1N)
id_cluster3_ren_mrg$CNTYGR1N[id_cluster3_ren_mrg $CNTYGR1N == "5"
   ] <- "Eastern␣Europe"
id_cluster3_ren_mrg$CNTYGR1N[id_cluster3_ren_mrg $CNTYGR1N == "6"
   ] <- "Western␣Europe␣and␣Israel"
id_cluster3_ren_mrg$CNTYGR1N[id_cluster3_ren_mrg $CNTYGR1N == "8"
   ] <- "North␣America"
id_cluster3_ren_mrg$BASEGFR <- as.character(id_cluster3_ren_mrg $
   BASEGFR)
id_cluster3_ren_mrg$LVEFHIM <- as.character(id_cluster3_ren_mrg $
   LVEFHIM)
id_cluster3_ren_mrg$AGEGR01 = as.factor(id_cluster3_ren_mrg$
   AGEGR01)
id_cluster3_ren_mrg$AGEGR02 = as.factor(id_cluster3_ren_mrg$
   AGEGR02)
id_cluster3_ren_mrg$SEX = as.factor(id_cluster3_ren_mrg$SEX)
id_cluster3_ren_mrg$RACE = as.factor(id_cluster3_ren_mrg$RACE)
id_cluster3_ren_mrg$CNTYGR1N = as.factor(id_cluster3_ren_mrg$
   CNTYGR1N)
id_cluster3_ren_mrg$NTBNP = as.factor(id_cluster3_ren_mrg$NTBNP)
id_cluster3_ren_mrg$NYHAFUC = as.factor(id_cluster3_ren_mrg$
   NYHAFUC)
```

```
id_cluster3_ren_mrg$BETABLK = as.factor(id_cluster3_ren_mrg$
   BETABLK)
id_cluster3_ren_mrg$PRIHOSP = as.factor(id_cluster3_ren_mrg$
   PRIHOSP)
id_cluster3_ren_mrg$DIABETE = as.factor(id_cluster3_ren_mrg$
   DIABETE)

id_cluster3_ren_mrg$ATRFIBR = as.factor(id_cluster3_ren_mrg$
   ATRFIBR)
id_cluster3_ren_mrg$HYPERTE = as.factor(id_cluster3_ren_mrg$
   HYPERTE)
id_cluster3_ren_mrg$MAXACE = as.factor(id_cluster3_ren_mrg$MAXACE
   )
id_cluster3_ren_mrg$MAXARB = as.factor(id_cluster3_ren_mrg$MAXARB
   )

id_cluster3_ren_mrg$BASEGFR = as.factor(id_cluster3_ren_mrg$
   BASEGFR)
id_cluster3_ren_mrg$BMIGR1 = as.factor(id_cluster3_ren_mrg$BMIGR1
   )
id_cluster3_ren_mrg$LVEFB25 = as.factor(id_cluster3_ren_mrg$
   LVEFB25)
id_cluster3_ren_mrg$LVEFHIM = as.factor(id_cluster3_ren_mrg$
   LVEFHIM)

names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "AGEGR01
   "] <- "Age_Group"
```

```
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "
   CNTYGR1N"] <- "Country_Group"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "NYHAFUC
   "] <- "NYHA"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "PRIHOSP
   "] <- "Prior_HF_hosp"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "DIABETE
   "] <- "Diabetes"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "ATRFIBR
   "] <- "Afib"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "HYPERTE
   "] <- "Hypertension"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "BASEGFR
   "] <- "EGFR"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "LVEFHIM
   "] <- "LVEF"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "BMIGR1"
   ] <- "BMI"
names(id_cluster3_ren_mrg)[names(id_cluster3_ren_mrg) == "V2"] <-
    "Cluster"

main_RSF_vars_cln <- id_cluster3_ren_mrg[!(id_cluster3_ren_mrg$
   NTBNP=="Missing" | id_cluster3_ren_mrg$EGFR==""),]
final_table <- table1(~ Activity_Intensity + Activity_Duration +
   factor(Age_Group) + factor(SEX) + factor(Country_Group) +
   factor(NTBNP) + factor(NYHA) + factor(Prior_HF_hosp) + factor(
   Diabetes) + factor(Afib) + factor(Hypertension) + factor(EGFR)
    + factor(BMI) | Cluster, data=main_RSF_vars_cln)
var_interest <- c("USUBJID", "Activity_Intensity", "Activity_
   Duration", "Age_Group", "SEX", "Country_Group", "NYHA", "NTBNP
   ", "Prior_HF_hosp", "Diabetes", "Afib", "Hypertension", "EGFR"
   , "BMI")
main_RSF_vars <- id_cluster3_ren_mrg[var_interest]
main_RSF_vars_cln <- main_RSF_vars[!(main_RSF_vars$NTBNP=="
   Missing" | main_RSF_vars$EGFR=="Missing"),]
main_RSF_vars <- main_RSF_vars_cln

clin_subset <- subset(adtte, PARAM=="HF␣HOSPITALIZATION" | PARAM
   =="HF␣HOSPITALIZATION␣AND␣URGENT␣VISIT␣FOR␣HF" | PARAM=="
   SECONDARY␣EFFICACY␣OUTCOME", select=c(USUBJID, PARAM, PARAMCD,
    AVAL, CNSR))
table(clin_subset$PARAM, clin_subset$CNSR)

clin_subset_hosp <- subset(adtte, PARAM=="HF␣HOSPITALIZATION" &
   TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣FIRST␣DOSE")
clin_subset_hosp_urg <- subset(adtte, PARAM=="HF␣HOSPITALIZATION␣
   AND␣URGENT␣VISIT␣FOR␣HF" & TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣
   FIRST␣DOSE")
clin_subset_secondary_eff <- subset(adtte, PARAM=="SECONDARY␣
   EFFICACY␣OUTCOME" & TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣FIRST␣DOSE"
   )
hosp_cnsr <- c("USUBJID", "AVAL", "CNSR")
clin_subset_hosp_cnsr <- clin_subset_hosp[hosp_cnsr]
```

```
RSF_hosp_final <- merge(main_RSF_vars, clin_subset_hosp_cnsr, by=
   "USUBJID")
RSF_hosp_final <- RSF_hosp_final[,-1] #Remove subject id since
   its not useful
RSF_hosp_final <- RSF_hosp_final[,-3] #Removing treatment
RSF_hosp_final$CNSR2 <- ifelse(RSF_hosp_final$CNSR==0, 1, 0)
RSF_hosp_final <- RSF_hosp_final[, -c(15)]

RSF_hosp_final_check0 <- subset(RSF_hosp_final, CNSR2==0)
RSF_hosp_final_check1 <- subset(RSF_hosp_final, CNSR2==1)
summary(RSF_hosp_final_check0$AVAL)
summary(RSF_hosp_final_check1$AVAL)

table(RSF_hosp_final$CNSR)

set.seed(12345)
Train <- createDataPartition(RSF_hosp_final$CNSR2, p=0.8, list=
   FALSE)
Training <- RSF_hosp_final[ Train, ]
Testing <- RSF_hosp_final[ -Train, ]
Training[sapply(Training, is.character)] <- lapply(Training[
   sapply(Training, is.character)],
                                       as.factor)
Testing[sapply(Testing, is.character)] <- lapply(Testing[sapply(
   Testing, is.character)],
                                      as.factor)
```

```
tab1(RSF_hosp_final$CNSR2, sort.group = "decreasing", cum.percent
    = TRUE) #85.9 and 14.1 after cleaning up the missing

options(rf.cores=20,mc.cores=20)
set.seed(12345)
nodesize<-c(10,20,35,50,70,85,100,120,150,180,190,200,210,220)
nsplit<-c(2,3,4,5,6,7,8,9,10,15,20)
mtry<-c(1,2,3,4,5,6,7,8,9,10,11, 12, 13, 14, 15)
combis<-expand.grid(nodesize,nsplit,mtry)
cv_time<-vector(mode="numeric",length=nrow(combis))
oob_error<-vector(mode="numeric",length=nrow(combis))
for(j in 1:nrow(combis)){
  cat("Working␣on␣combination␣n␣r:",j,"with␣nodesize:",
     combis[j,1],
     "nsplit:",combis[j,2],"and␣mtry:",
     combis[j,3],"\n")
  cv_time[j]<-{
   system.time(
     fit_ovr<-rfsrc(Surv(AVAL,CNSR2)~.,
                 splitrule="logrank",
                 nodesize=combis[j,1],
                 nsplit=combis[j,2],
                 data=Training,
                 mtry=combis[j,3],
                 ntree=500,seed=-12345,
                 # sampsize=3473,
                 ntime=500,forest=FALSE
    ))[3]
  }
  cat("Calculating␣the␣OOB␣prediction␣error␣of␣combination...",
    "\n")
  oob_error[j]<-as.numeric(fit_ovr$err.rate[fit_ovr$ntree])
  cat("Error␣is:",oob_error[j],"\n")
}
df_logrank_ovr<-data.frame(Node_size=combis$Var1,
                    Nsplit=combis$Var2,
                    Mtry=combis$Var3,
                    Error=round(oob_error,4))
df_logrank_ovr[which.min(df_logrank_ovr$Error),]

bestnode=df_logrank_ovr[which.min(df_logrank_ovr$Error),][1]
bestnsplit=df_logrank_ovr[which.min(df_logrank_ovr$Error),][2]
bestmtry=df_logrank_ovr[which.min(df_logrank_ovr$Error),][3]

fit<-rfsrc(Surv(AVAL,CNSR2)~.,
        splitrule="logrank",nsplit=2,
        data=Training,ntree=500,
        split.depth="all.trees",
        var.used="all.trees",seed=-12345,
        mtry=4,nodesize=20,#nodedepths=7,
        ntime=500,forest=TRUE,
        importance=TRUE
)
```

```
fit$err.rate[fit$ntree]


plot(gg_vimp(fit), labs = st.labs)+
  theme(legend.position = c(0.8, 0.2)) +
  labs(fill = "VIMP␣>␣0")


summary(fit$time.interest)


gg_v <- gg_variable(fit, time = c(6, 176),
                    time.labels = c("1␣day", "6␣months"))


plot(gg_v, xvar = "NTBNP", alpha = 0.4)
plot(gg_v, xvar = "Age_Group", alpha = 0.4)
plot(gg_v, xvar = "SEX", alpha = 0.4)
plot(gg_v, xvar = "Country_Group", alpha = 0.4)
plot(gg_v, xvar = "NYHA", alpha = 0.4)
plot(gg_v, xvar = "Prior_HF_hosp", alpha = 0.4)
plot(gg_v, xvar = "Diabetes", alpha = 0.4)
plot(gg_v, xvar = "Afib", alpha = 0.4)
plot(gg_v, xvar = "Hypertension", alpha = 0.4)
plot(gg_v, xvar = "EGFR", alpha = 0.4)
plot(gg_v, xvar = "BMI", alpha = 0.4)
plot(gg_v, xvar = "Activity_Intensity", alpha = 0.4)
plot(gg_v, xvar = "Activity_Duration", alpha = 0.4)



varsel_pbc <- var.select(fit)
```

```
gg_md <- gg_minimal_depth(varsel_pbc, lbls = st.labs)
plot(gg_md)
print(gg_md)

plot(gg_minimal_vimp(gg_md))# , lbls = st.labs)
theme(legend.position=c(0.8, 0.2))

RSF_hosp_final_removed <- RSF_hosp_final[, -c(3, 11, 13)]

set.seed(12345)
Train <- createDataPartition(RSF_hosp_final_removed$CNSR2, p=0.8,
    list=FALSE)
Training <- RSF_hosp_final_removed[ Train, ]
Testing <- RSF_hosp_final_removed[ -Train, ]
Training[sapply(Training, is.character)] <- lapply(Training[
   sapply(Training, is.character)],
                                       as.factor)
Testing[sapply(Testing, is.character)] <- lapply(Testing[sapply(
   Testing, is.character)],
                                     as.factor)
tab1(RSF_hosp_final_removed$CNSR2, sort.group = "decreasing", cum
   .percent = TRUE) #85.9 vs. 14.4%

options(rf.cores=20,mc.cores=20)
set.seed(12345)
nodesize<-c(10,20,35,50,70,85,100,120,150,180,190,200,210,220)
nsplit<-c(2,3,4,5,6,7,8,9,10,15,20)
```

```
mtry<-c(1,2,3,4,5,6,7,8,9,10,15,20,35,40,50)
combis<-expand.grid(nodesize,nsplit,mtry)
cv_time<-vector(mode="numeric",length=nrow(combis))
oob_error<-vector(mode="numeric",length=nrow(combis))
for(j in 1:nrow(combis)){
  cat("Working␣on␣combination␣n␣r:",j,"with␣nodesize:",
      combis[j,1],
      "nsplit:",combis[j,2],"and␣mtry:",
      combis[j,3],"\n")
  cv_time[j]<-{
    system.time(
      fit_ovr<-rfsrc(Surv(AVAL,CNSR2)~.,
                     splitrule="logrank",
                     nodesize=combis[j,1],
                     nsplit=combis[j,2],
                     data=Training,
                     mtry=combis[j,3],
                     ntree=500,seed=-12345,
                     # sampsize=3473,
                     ntime=500,forest=FALSE
      ))[3]
  }
  cat("Calculating␣the␣OOB␣prediction␣error␣of␣combination...",
      "\n")
  oob_error[j]<-as.numeric(fit_ovr$err.rate[fit_ovr$ntree])
  cat("Error␣is:",oob_error[j],"\n")
}
```

```
df_logrank_ovr<-data.frame(Node_size=combis$Var1,
                    Nsplit=combis$Var2,
                    Mtry=combis$Var3,
                    Error=round(oob_error,4))
df_logrank_ovr[which.min(df_logrank_ovr$Error),]

bestnode=df_logrank_ovr[which.min(df_logrank_ovr$Error),][1]
bestnsplit=df_logrank_ovr[which.min(df_logrank_ovr$Error),][2]
bestmtry=df_logrank_ovr[which.min(df_logrank_ovr$Error),][3]
fit<-rfsrc(Surv(AVAL,CNSR2)~.,
        splitrule="logrank",nsplit=3,
        data=Training,ntree=500,
        split.depth="all.trees",
        var.used="all.trees",seed=-12345,
        mtry=1,nodesize=20,#nodedepths=7,
        ntime=500,forest=TRUE,
        importance=TRUE
)

fit$err.rate[fit$ntree]

plot(gg_vimp(fit), labs = st.labs)+
  theme(legend.position = c(0.8, 0.2)) +
  labs(fill = "VIMP␣>␣0")

summary(fit$time.interest)
```

```
gg_v <- gg_variable(fit, time = c(6, 176),
               time.labels = c("1␣day", "6␣months"))

plot(gg_v, xvar = "NTBNP", alpha = 0.4)
plot(gg_v, xvar = "Age_Group", alpha = 0.4)
plot(gg_v, xvar = "SEX", alpha = 0.4)
plot(gg_v, xvar = "Country_Group", alpha = 0.4)
plot(gg_v, xvar = "NYHA", alpha = 0.4)
plot(gg_v, xvar = "Prior_HF_hosp", alpha = 0.4)
plot(gg_v, xvar = "Diabetes", alpha = 0.4)
plot(gg_v, xvar = "Afib", alpha = 0.4)
plot(gg_v, xvar = "Hypertension", alpha = 0.4)
plot(gg_v, xvar = "EGFR", alpha = 0.4)
plot(gg_v, xvar = "BMI", alpha = 0.4)
plot(gg_v, xvar = "Activity_Intensity", alpha = 0.4)
plot(gg_v, xvar = "Activity_Duration", alpha = 0.4)

varsel_pbc <- var.select(fit)
gg_md <- gg_minimal_depth(varsel_pbc, lbls = st.labs)
plot(gg_md)
print(gg_md)

plot(gg_minimal_vimp(gg_md))# , lbls = st.labs)
theme(legend.position=c(0.8, 0.2))

hosp_urg_cnsr <- c("USUBJID", "AVAL", "CNSR")
clin_subset_hosp_urg_cnsr <- clin_subset_hosp_urg[hosp_urg_cnsr]
```

```
RSF_hosp_urg_final <- merge(main_RSF_vars, clin_subset_hosp_urg_
   cnsr, by="USUBJID")
RSF_hosp_urg_final <- RSF_hosp_urg_final[,-1]
RSF_hosp_urg_final <- RSF_hosp_urg_final[,-3]
RSF_hosp_urg_final$CNSR2 <- ifelse(RSF_hosp_urg_final$CNSR==0, 1,
    0)
RSF_hosp_urg_final <- RSF_hosp_urg_final[, -c(16)]
RSF_hosp_urg_final <- RSF_hosp_urg_final[, -c(15)]
names(RSF_hosp_urg_final)

RSF_hosp_urg_final_check0 <- subset(RSF_hosp_urg_final, CNSR2==0)
RSF_hosp_urg_final_check1 <- subset(RSF_hosp_urg_final, CNSR2==1)
summary(RSF_hosp_urg_final_check0$AVAL)
summary(RSF_hosp_urg_final_check1$AVAL)
set.seed(12345)
Train <- createDataPartition(RSF_hosp_urg_final$CNSR2, p=0.8,
   list=FALSE)
Training <- RSF_hosp_urg_final[ Train, ]
Testing <- RSF_hosp_urg_final[ -Train, ]
Training[sapply(Training, is.character)] <- lapply(Training[
   sapply(Training, is.character)],
                                       as.factor)
Testing[sapply(Testing, is.character)] <- lapply(Testing[sapply(
   Testing, is.character)],
                                     as.factor)
```

```
tab1(RSF_hosp_urg_final$CNSR2, sort.group = "decreasing", cum.
   percent = TRUE) #84.1% vs 15.9%

options(rf.cores=20,mc.cores=20)
set.seed(12345)
nodesize<-c(10,20,35,50,70,85,100,120,150,180,190,200,210,220)
nsplit<-c(2,3,4,5,6,7,8,9,10,15,20)
mtry<-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)
combis<-expand.grid(nodesize,nsplit,mtry)
cv_time<-vector(mode="numeric",length=nrow(combis))
oob_error<-vector(mode="numeric",length=nrow(combis))
for(j in 1:nrow(combis)){
  cat("Working␣on␣combination␣n␣r:",j,"with␣nodesize:",
     combis[j,1],
     "nsplit:",combis[j,2],"and␣mtry:",
     combis[j,3],"\n")
  cv_time[j]<-{
   system.time(
     fit_ovr<-rfsrc(Surv(AVAL,CNSR2)~.,
                 splitrule="logrank",
                 nodesize=combis[j,1],
                 nsplit=combis[j,2],
                 data=Training,
                 mtry=combis[j,3],
                 ntree=500,seed=-12345,
                 # sampsize=3473,
                 ntime=500,forest=FALSE
```

```
    ))[3]
  }
  cat("Calculating␣the␣OOB␣prediction␣error␣of␣combination...",
    "\n")
  oob_error[j]<-as.numeric(fit_ovr$err.rate[fit_ovr$ntree])
  cat("Error␣is:",oob_error[j],"\n")
}
df_logrank_ovr<-data.frame(Node_size=combis$Var1,
                    Nsplit=combis$Var2,
                    Mtry=combis$Var3,
                    Error=round(oob_error,4))
df_logrank_ovr[which.min(df_logrank_ovr$Error),]

bestnode=df_logrank_ovr[which.min(df_logrank_ovr$Error),][1]
bestnsplit=df_logrank_ovr[which.min(df_logrank_ovr$Error),][2]
bestmtry=df_logrank_ovr[which.min(df_logrank_ovr$Error),][3]

fit<-rfsrc(Surv(AVAL,CNSR2)~.,
        splitrule="logrank",nsplit=2,
        data=Training,ntree=500,
        split.depth="all.trees",
        var.used="all.trees",seed=-12345,
        mtry=4,nodesize=20,#nodedepths=7,
        ntime=500,forest=TRUE,
        importance=TRUE
)
```

```
fit$err.rate[fit$ntree]

plot(gg_vimp(fit), labs = st.labs)+
  theme(legend.position = c(0.8, 0.2)) +
  labs(fill = "VIMP␣>␣0")

summary(fit$time.interest)

gg_v <- gg_variable(fit, time = c(8, 176),
               time.labels = c("1␣day", "6␣months"))

plot(gg_v, xvar = "NTBNP", alpha = 0.4)
plot(gg_v, xvar = "Age_Group", alpha = 0.4)
plot(gg_v, xvar = "SEX", alpha = 0.4)
plot(gg_v, xvar = "Country_Group", alpha = 0.4)
plot(gg_v, xvar = "NYHA", alpha = 0.4)
plot(gg_v, xvar = "Prior_HF_hosp", alpha = 0.4)
plot(gg_v, xvar = "Diabetes", alpha = 0.4)
plot(gg_v, xvar = "Afib", alpha = 0.4)
plot(gg_v, xvar = "Hypertension", alpha = 0.4)
plot(gg_v, xvar = "EGFR", alpha = 0.4)
plot(gg_v, xvar = "BMI", alpha = 0.4)
plot(gg_v, xvar = "Activity_Intensity", alpha = 0.4)
plot(gg_v, xvar = "Activity_Duration", alpha = 0.4)

varsel_pbc <- var.select(fit)
gg_md <- gg_minimal_depth(varsel_pbc, lbls = st.labs)
```

```
plot(gg_md)
print(gg_md)

hosp_urg_cnsr <- c("USUBJID", "AVAL", "CNSR")
clin_subset_hosp_urg_cnsr <- clin_subset_hosp_urg[hosp_urg_cnsr]

RSF_hosp_urg_final_v2 <- merge(main_RSF_vars, clin_subset_hosp_
   urg_cnsr, by="USUBJID")
RSF_hosp_urg_final_v2 <- RSF_hosp_urg_final_v2[,-c(1, 4, 6, 7,
   13, 14, 15, 18)]
RSF_hosp_urg_final_v2$CNSR2 <- ifelse(RSF_hosp_urg_final_v2$CNSR
   ==0, 1, 0)
RSF_hosp_urg_final_v2 <- RSF_hosp_urg_final_v2[, -c(10)]

names(RSF_hosp_urg_final_v2)

RSF_hosp_urg_final_v2_check0 <- subset(RSF_hosp_urg_final_v2,
   CNSR2==0)
RSF_hosp_urg_final_v2_check1 <- subset(RSF_hosp_urg_final_v2,
   CNSR2==1)
summary(RSF_hosp_urg_final_v2_check0$AVAL)
summary(RSF_hosp_urg_final_v2_check1$AVAL)
set.seed(12345)
Train <- createDataPartition(RSF_hosp_urg_final_v2$CNSR2, p=0.8,
   list=FALSE)
Training <- RSF_hosp_urg_final_v2[ Train, ]
Testing <- RSF_hosp_urg_final_v2[ -Train, ]
```

```
Training[sapply(Training, is.character)] <- lapply(Training[
   sapply(Training, is.character)],
                                                  as.factor)
Testing[sapply(Testing, is.character)] <- lapply(Testing[sapply(
   Testing, is.character)],
                                                as.factor)
tab1(RSF_hosp_urg_final_v2$CNSR2, sort.group = "decreasing", cum.
   percent = TRUE) #84.1% vs 15.9%
options(rf.cores=20,mc.cores=20)

set.seed(12345)
nodesize<-c(10,20,35,50,70,85,100,120,150,180,190,200,210,220)
nsplit<-c(2,3,4,5,6,7,8,9,10,15,20)
mtry<-c(1,2,3,4,5,6,7,8,9,10,15,20,35,40,50)
combis<-expand.grid(nodesize,nsplit,mtry)
cv_time<-vector(mode="numeric",length=nrow(combis))
oob_error<-vector(mode="numeric",length=nrow(combis))
for(j in 1:nrow(combis)){
  cat("Working␣on␣combination␣n␣r:",j,"with␣nodesize:",
      combis[j,1],
      "nsplit:",combis[j,2],"and␣mtry:",
      combis[j,3],"\n")
  cv_time[j]<-{
    system.time(
      fit_ovr<-rfsrc(Surv(AVAL,CNSR2)~.,
                  splitrule="logrank",
                  nodesize=combis[j,1],
                nsplit=combis[j,2],
                data=Training,
                mtry=combis[j,3],
                ntree=500,seed=-12345,
                # sampsize=3473,
                ntime=500,forest=FALSE
    ))[3]
  }
  cat("Calculating␣the␣OOB␣prediction␣error␣of␣combination...",
    "\n")
  oob_error[j]<-as.numeric(fit_ovr$err.rate[fit_ovr$ntree])
  cat("Error␣is:",oob_error[j],"\n")
}
df_logrank_ovr<-data.frame(Node_size=combis$Var1,
                    Nsplit=combis$Var2,
                    Mtry=combis$Var3,
                    Error=round(oob_error,4))
df_logrank_ovr[which.min(df_logrank_ovr$Error),]

bestnode=df_logrank_ovr[which.min(df_logrank_ovr$Error),][1]
bestnsplit=df_logrank_ovr[which.min(df_logrank_ovr$Error),][2]
bestmtry=df_logrank_ovr[which.min(df_logrank_ovr$Error),][3]

fit<-rfsrc(Surv(AVAL,CNSR2)~.,
        splitrule="logrank",nsplit=4,
        data=Training,ntree=500,
        split.depth="all.trees",
        var.used="all.trees",seed=-12345,
        mtry=8,nodesize=10,#nodedepths=7,
        ntime=500,forest=TRUE,
        importance=TRUE
)

fit$err.rate[fit$ntree]

plot(gg_vimp(fit), labs = st.labs)+
  theme(legend.position = c(0.8, 0.2)) +
  labs(fill = "VIMP␣>␣0")

summary(fit$time.interest)

gg_v <- gg_variable(fit, time = c(6, 176),
               time.labels = c("1␣day", "6␣months"))

plot(gg_v, xvar = "NTBNP", alpha = 0.4)
plot(gg_v, xvar = "Age_Group", alpha = 0.4)
plot(gg_v, xvar = "SEX", alpha = 0.4)
plot(gg_v, xvar = "Country_Group", alpha = 0.4)
plot(gg_v, xvar = "NYHA", alpha = 0.4)
plot(gg_v, xvar = "Prior_HF_hosp", alpha = 0.4)
plot(gg_v, xvar = "Diabetes", alpha = 0.4)
plot(gg_v, xvar = "Afib", alpha = 0.4)
plot(gg_v, xvar = "Hypertension", alpha = 0.4)
plot(gg_v, xvar = "EGFR", alpha = 0.4)
plot(gg_v, xvar = "BMI", alpha = 0.4)
plot(gg_v, xvar = "Activity_Intensity", alpha = 0.4)
plot(gg_v, xvar = "Activity_Duration", alpha = 0.4)

varsel_pbc <- var.select(fit)
gg_md <- gg_minimal_depth(varsel_pbc, lbls = st.labs)
print(gg_md)

plot(gg_minimal_vimp(gg_md))# , lbls = st.labs)
theme(legend.position=c(0.8, 0.2))

clin_subset_secondary_eff <- subset(adtte, PARAM=="SECONDARY␣
   EFFICACY␣OUTCOME" & TIMEREF=="UP␣TO␣26␣WEEKS␣AFTER␣FIRST␣DOSE"
   )

secondary_cnsr <- c("USUBJID", "AVAL", "CNSR")
clin_subset_sec_cnsr <- clin_subset_secondary_eff[secondary_cnsr]

RSF_sec_final <- merge(main_RSF_vars, clin_subset_sec_cnsr, by="
   USUBJID")
RSF_sec_final <- RSF_sec_final[,-c(1, 4, 18)]
RSF_sec_final$CNSR2 <- ifelse(RSF_sec_final$CNSR==0, 1, 0)
table(RSF_sec_final$CNSR)
RSF_sec_final <- RSF_sec_final[, -c(15)]
names(RSF_sec_final)

RSF_sec_final0 <- subset(RSF_sec_final, CNSR2==0)
```

```
RSF_sec_final1 <- subset(RSF_sec_final, CNSR2==1)
summary(RSF_sec_final0$AVAL)
summary(RSF_sec_final1$AVAL)

set.seed(12345)
Train <- createDataPartition(RSF_sec_final$CNSR2, p=0.8, list=
   FALSE)
Training <- RSF_sec_final[ Train, ]
Testing <- RSF_sec_final[ -Train, ]
Training[sapply(Training, is.character)] <- lapply(Training[
   sapply(Training, is.character)],
                                      as.factor)
Testing[sapply(Testing, is.character)] <- lapply(Testing[sapply(
   Testing, is.character)],
                                     as.factor)
tab1(RSF_sec_final$CNSR2, sort.group = "decreasing", cum.percent
   = TRUE) #83.3% vs 16.7%

options(rf.cores=20,mc.cores=20)


set.seed(12345)
nodesize<-c(10,20,35,50,70,85,100,120,150,180,190,200,210,220)
nsplit<-c(2,3,4,5,6,7,8,9,10,15,20)
mtry<-c(1,2,3,4,5,6,7,8,9,10,15,20,35,40,50)
combis<-expand.grid(nodesize,nsplit,mtry)
cv_time<-vector(mode="numeric",length=nrow(combis))
oob_error<-vector(mode="numeric",length=nrow(combis))
```

```
for(j in 1:nrow(combis)){
  cat("Working␣on␣combination␣n␣r:",j,"with␣nodesize:",
     combis[j,1],
     "nsplit:",combis[j,2],"and␣mtry:",
     combis[j,3],"\n")
  cv_time[j]<-{
    system.time(
     fit_ovr<-rfsrc(Surv(AVAL,CNSR2)~.,
                 splitrule="logrank",
                 nodesize=combis[j,1],
                 nsplit=combis[j,2],
                 data=Training,
                 mtry=combis[j,3],
                 ntree=500,seed=-12345,
                 # sampsize=3473,
                 ntime=500,forest=FALSE
     ))[3]
  }
  cat("Calculating␣the␣OOB␣prediction␣error␣of␣combination...",
     "\n")
  oob_error[j]<-as.numeric(fit_ovr$err.rate[fit_ovr$ntree])
  cat("Error␣is:",oob_error[j],"\n")
}
df_logrank_ovr<-data.frame(Node_size=combis$Var1,
                      Nsplit=combis$Var2,
                      Mtry=combis$Var3,
                      Error=round(oob_error,4))
```

```
df_logrank_ovr[which.min(df_logrank_ovr$Error),]


bestnode=df_logrank_ovr[which.min(df_logrank_ovr$Error),][1]
bestnsplit=df_logrank_ovr[which.min(df_logrank_ovr$Error),][2]
bestmtry=df_logrank_ovr[which.min(df_logrank_ovr$Error),][3]


fit<-rfsrc(Surv(AVAL,CNSR2)~.,
        splitrule="logrank",nsplit=2,
        data=Training,ntree=500,
        split.depth="all.trees",
        var.used="all.trees",seed=-12345,
        mtry=3,nodesize=35,#nodedepths=7,
        ntime=500,forest=TRUE,
        importance=TRUE
)


fit$err.rate[fit$ntree]


plot(gg_vimp(fit), labs = st.labs)+
  theme(legend.position = c(0.8, 0.2)) +
  labs(fill = "VIMP␣>␣0")


summary(fit$time.interest)


gg_v <- gg_variable(fit, time = c(6, 176),
               time.labels = c("1␣day", "6␣months"))
```

```
plot(gg_v, xvar = "NTBNP", alpha = 0.4)
plot(gg_v, xvar = "Age_Group", alpha = 0.4)
plot(gg_v, xvar = "SEX", alpha = 0.4)
plot(gg_v, xvar = "Country_Group", alpha = 0.4)
plot(gg_v, xvar = "NYHA", alpha = 0.4)
plot(gg_v, xvar = "Prior_HF_hosp", alpha = 0.4)
plot(gg_v, xvar = "Diabetes", alpha = 0.4)
plot(gg_v, xvar = "Afib", alpha = 0.4)
plot(gg_v, xvar = "Hypertension", alpha = 0.4)
plot(gg_v, xvar = "EGFR", alpha = 0.4)
plot(gg_v, xvar = "BMI", alpha = 0.4)
plot(gg_v, xvar = "Activity_Intensity", alpha = 0.4)
plot(gg_v, xvar = "Activity_Duration", alpha = 0.4)

varsel_pbc <- var.select(fit)
gg_md <- gg_minimal_depth(varsel_pbc, lbls = st.labs)
print(gg_md)
plot(gg_md)

plot(gg_minimal_vimp(gg_md))# , lbls = st.labs)
theme(legend.position=c(0.8, 0.2))

RSF_sec_final_rem <- RSF_sec_final[,-c(3, 5, 11)]

set.seed(12345)
Train <- createDataPartition(RSF_sec_final_rem$CNSR2, p=0.8, list
   =FALSE)
```

```
Training <- RSF_sec_final_rem[ Train, ]
Testing <- RSF_sec_final_rem[ -Train, ]
Training[sapply(Training, is.character)] <- lapply(Training[
   sapply(Training, is.character)],
                                          as.factor)
Testing[sapply(Testing, is.character)] <- lapply(Testing[sapply(
   Testing, is.character)],
                                        as.factor)
tab1(RSF_sec_final_rem$CNSR2, sort.group = "decreasing", cum.
   percent = TRUE) #83% vs 17%

options(rf.cores=20,mc.cores=20)

set.seed(12345)
nodesize<-c(10,20,35,50,70,85,100,120,150,180,190,200,210,220)
nsplit<-c(2,3,4,5,6,7,8,9,10,15,20)
mtry<-c(1,2,3,4,5,6,7,8,9,10,15,20,35,40,50)
combis<-expand.grid(nodesize,nsplit,mtry)
cv_time<-vector(mode="numeric",length=nrow(combis))
oob_error<-vector(mode="numeric",length=nrow(combis))
for(j in 1:nrow(combis)){
  cat("Working␣on␣combination␣n␣r:",j,"with␣nodesize:",
     combis[j,1],
     "nsplit:",combis[j,2],"and␣mtry:",
     combis[j,3],"\n")
  cv_time[j]<-{
   system.time(
    fit_ovr<-rfsrc(Surv(AVAL,CNSR2)~.,
               splitrule="logrank",
               nodesize=combis[j,1],
               nsplit=combis[j,2],
               data=Training,
               mtry=combis[j,3],
               ntree=500,seed=-12345,
               # sampsize=3473,
               ntime=500,forest=FALSE
    ))[3]
  }
  cat("Calculating␣the␣OOB␣prediction␣error␣of␣combination...",
    "\n")
  oob_error[j]<-as.numeric(fit_ovr$err.rate[fit_ovr$ntree])
  cat("Error␣is:",oob_error[j],"\n")
}
df_logrank_ovr<-data.frame(Node_size=combis$Var1,
                    Nsplit=combis$Var2,
                    Mtry=combis$Var3,
                    Error=round(oob_error,4))
df_logrank_ovr[which.min(df_logrank_ovr$Error),]

bestnode=df_logrank_ovr[which.min(df_logrank_ovr$Error),][1]
bestnsplit=df_logrank_ovr[which.min(df_logrank_ovr$Error),][2]
bestmtry=df_logrank_ovr[which.min(df_logrank_ovr$Error),][3]

fit_v2<-rfsrc(Surv(AVAL,CNSR2)~.,
           splitrule="logrank",nsplit=3,
           data=Training,ntree=500,
           split.depth="all.trees",
           var.used="all.trees",seed=-12345,
           mtry=2,nodesize=20,#nodedepths=7,
           ntime=500,forest=TRUE,
           importance=TRUE
)

fit_v2$err.rate[fit_v2$ntree]

plot(gg_vimp(fit_v2), labs = st.labs)+
  theme(legend.position = c(0.8, 0.2)) +
  labs(fill = "VIMP␣>␣0")
RSF_sec_final_rem <- RSF_sec_final[,-c(3, 5, 11, 12, 13)]

fit_v2<-rfsrc(Surv(AVAL,CNSR2)~.,
        splitrule="logrank",nsplit=2,
        data=Training,ntree=500,
        split.depth="all.trees",
        var.used="all.trees",seed=-12345,
        mtry=2,nodesize=35,#nodedepths=7,
        ntime=500,forest=TRUE,
        importance=TRUE
)

fit_v2$err.rate[fit_v2$ntree]
```

```
plot(gg_vimp(fit_v2), labs = st.labs)+
 theme(legend.position = c(0.8, 0.2)) +
 labs(fill = "VIMP␣>␣0")

summary(fit_v2$time.interest)

gg_v <- gg_variable(fit_v2, time = c(6, 176),
               time.labels = c("1␣day", "6␣months"))

plot(gg_v, xvar = "NTBNP", alpha = 0.4)
plot(gg_v, xvar = "Age_Group", alpha = 0.4)
plot(gg_v, xvar = "NYHA", alpha = 0.4)
plot(gg_v, xvar = "Diabetes", alpha = 0.4)
plot(gg_v, xvar = "Afib", alpha = 0.4)
plot(gg_v, xvar = "EGFR", alpha = 0.4)
plot(gg_v, xvar = "BMI", alpha = 0.4)
plot(gg_v, xvar = "Activity_Intensity", alpha = 0.4)
plot(gg_v, xvar = "Activity_Duration", alpha = 0.4)

varsel_pbc <- var.select(fit_v2)
gg_md <- gg_minimal_depth(varsel_pbc, lbls = st.labs)
print(gg_md)

plot(gg_minimal_vimp(gg_md))# , lbls = st.labs)
theme(legend.position=c(0.8, 0.2))
```