Exploring the potential of routine cancer data

Raghul Seerangan
201686442

Supervised by:
Duncan Wilson, Kara-Lousie Royle, Lesley Smith (Leeds Institute of Clinical Trials Research)

Submitted in accordance with the requirements for the degree of of

Master of Science in Data Science and Analystics

The University of Leeds, School of Mathematics

August 2023

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

width=!,height=!,

Abstract

Multiple myeloma, a malignancy originating from the plasma cells in the bone marrow, presents a significant health challenge. This study aimed to elucidate the various risk factors associated with this condition using a dataset of 12,492 observations and 62 variables. Key variables such as age, gender, and comorbidity scores were analyzed, revealing males represented 57% of the population and the age distribution was predominantly between 40 to 95 years. Using the Cox Proportional Hazards Model, 'Age', 'Time To First Treatment', and comorbidity scores emerged as significant predictors for survival, with a concordance index of 0.668. Additionally, most patients who succumbed had lower comorbidity scores. This research provides a comprehensive understanding of the factors affecting survival rates in multiple myeloma patients, offering valuable insights for healthcare practitioners and researchers.

Contents

1 Introduction **1**

1.1 Background and Motivation 1

1.1.1 Background of Multiple Myeloma and Treatment Lines 1

1.1.2 The Gold Standard in Clinical Trials and The Role of Routine Data . . 1

1.1.3 Simulacrum: A Simulated Data Beacon 2

1.1.4 Dataset Context 2

1.1.5 Simulacrum Unveiled 2

1.2 Aim and Objectives of the Study 3

1.2.1 Aim 3

1.2.2 Objectives 3

2 Literature Review **5**

2.1 Historical Overview of Cancer Data Collection 5

2.1.1 Pre-20th Century 5

2.1.2 The Birth of Cancer Registries (20th Century) 5

2.1.3 The Advent of Digital Data Collection 5

2.1.4 Genomic Data and Personalized Medicine 6

2.1.5 Big Data and the Future of Cancer Data Collection 6

2.2 Previous Studies Leveraging Routine Cancer Data 6

2.2.1 Epidemiological Insights 6

2.2.2 Treatment Outcomes and Effectiveness 7

2.2.3 Health Disparities and Socio-economic Factors 7

2.2.4 Predictive Models and Machine Learning 7

2.3 The Role of Statistical Models in Oncological Research 7

2.3.1 The Rise of Statistical Models 7

2.3.2 Quantifying Uncertainties 8

2.3.3 Personalized Medicine and Predictive Modelling 8

2.3.4 Facilitating Clinical Trials 8

2.3.5 Challenges and the Road Ahead 8

2.4 Gaps in Current Literature and the Need for the Current Study 9

2.4.1 The Ever-Evolving Landscape of Oncology Research 9

2.4.2 Routine Cancer Data - A Treasure Trove Often Overlooked 9

2.4.3 Statistical Models - Depth Over Breadth 9

2.4.4 The Interplay Between Clinical and Computational Research 9

3 Methodology **11**
3.1 Data Source and Collection . . . 11
3.1.1 Origins of the Data . . . 11
3.1.2 Data Collection Mechanism . . . 12
3.1.3 Selection Criteria and Data Extraction . . . 12
3.1.4 Data Complexity and Connections . . . 14
3.1.5 Data Description . . . 15
3.2 Data Preprocessing . . . 18
3.2.1 Merging Data Across Tables . . . 18
3.2.2 Filtering Relevant Records . . . 19
3.2.3 Column Removal . . . 21
3.2.4 Handling Missing Values . . . 22
3.3 Descriptive analysis . . . 24
3.3.1 Understanding the dataset . . . 24
3.3.2 CATEGORICAL FEATURES . . . 24
3.3.3 Age Distribution Analysis . . . 26
3.3.4 Distribution of Body Mass Index (BMI) Among Patients . . . 27
3.3.5 Dead Event . . . 27
3.3.6 Data Correlation . . . 28
3.4 Survival analysis . . . 29
3.4.1 Survival Function . . . 30
3.4.2 Estimated via Kaplan-Meier . . . 31
3.4.3 Interpreting the Kaplan-Meier Survival Curve . . . 31
3.5 Cox Proportional Hazards Model . . . 32
3.5.1 Hazard and Hazard Ratio . . . 33
3.5.2 Assumption of Proportional Hazards . . . 33

4 Results and Discussion **35**
4.1 Key Determinants of Survival Rates . . . 35
4.1.1 The Impact of Age on Survival . . . 35
4.1.2 Gender and Survival Rate . . . 36
4.1.3 Clinical Trail Survival Rate . . . 36
4.1.4 Comorbidity Score and its effects on the survival rate . . . 37
4.2 Survival Analysis by Different Groups . . . 38
4.2.1 Survival Analysis by Age . . . 38
4.2.2 Survival Analysis by Comorbidity Score . . . 39
4.2.3 Survival Analysis by BMI . . . 41
4.2.4 Survival Analysis by Time to First Treatment . . . 41
4.2.5 Survival Analysis by Cancer Care Plan Intent . . . 42
4.2.6 Survival Analysis by Gender . . . 43
4.3 Cox Proportional Hazards Model . . . 44
4.3.1 Cox PH Fitter . . . 44
4.3.2 Goodness of Fit . . . 44
4.3.3 Statistical Significance and Its Importance . . . 46
4.3.4 Assessing the Predictive Strength of Variables in Routine Multiple Myeloma Cancer Data . . . 47
4.3.5 Harnessing the Power of Routine Data in Multiple Myeloma Patient Outcomes . . . 48

4.3.6 Deciphering Predictive Potency through Concordance in the Realm of Multiple Myeloma Data 49
4.3.7 Partial Akaike Information Criterion (AIC) in the Landscape of Multiple Myeloma Data 49
4.3.8 Understanding the Partial Effects of Covariate Roles in Multiple Myeloma Survival Stories 50
4.3.9 The Illuminating Power of the Log-Likelihood Ratio Test 53

5 Conclusion **55**
5.1 From Raw Data to Revelations: A Myeloma Exploration 55
5.2 Key Takeaways 56
5.2.1 Age as a Silent Protagonist 56
5.2.2 The Gendered Dance of Survival 56
5.2.3 Clinical Trials as Beacons 56
5.2.4 The Weight of Comorbidities 56
5.2.5 Timeliness in Treatment 57
5.3 Implications for Clinical Practice 57
5.4 Future Directions 58
5.5 Limitations 59

List of Figures

3.1 Categorical features . . . 25
3.2 Distribution of age . . . 26
3.3 Age Distribution w.r.t Clinical Trial . . . 27
3.4 Distribution of BMI Among Patients . . . 27
3.5 Distribution of death event . . . 28
3.6 Heatmap . . . 29

4.1 Effect of age on survival . . . 35
4.2 Survive rate w.r.t gender . . . 36
4.3 Clinical Trail Survival Rate . . . 37
4.4 Effects on the survival rate . . . 37
4.5 Distribution of Comorbidity Score . . . 38
4.6 Kaplan Meier Survival Curve by Age Groups . . . 38
4.7 Kaplan Meier Survival Curve by Comorbidity Score . . . 40
4.8 Kaplan Meier Survival curve by BMI . . . 41
4.9 Kaplan Meier Survival curve by Time to First Treatment . . . 42
4.10 Kaplan Meier Survival curve by Cancer Care Plan Intent . . . 42
4.11 Kaplan Meier Survival curve by Gender . . . 43
4.12 Summary of Coxfitter . . . 44
4.13 Visual representation of the log hazard ratios . . . 45
4.14 Visual representation of the hazard ratios . . . 45
4.15 Result of Cox Model . . . 48
4.16 Partial Effects of Varied Time To first treatment . . . 50
4.17 Partial Effects of Varied Age . . . 51
4.18 Partial Effects of Varied Comorbidity . . . 52
4.19 Partial Effects of Varied PERF STATUS START OF CYCLE . . . 53

List of Tables

3.1 Dataset Column Descriptions 15

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Background of Multiple Myeloma and Treatment Lines

Multiple myeloma is a common type of blood cancer that comes back over and over again. This cancer is hard to treat because it tends to get worse even when it seems to be under control for a while. As a person with multiple myeloma fights the disease, they will encounter different kinds of treatment, each of which is called a ”line of treatment.” Some treatments have a set number of cycles that must be completed, while others may be stopped if they cause severe side effects. The Leeds Institute of Clinical Trials Research wants to improve the outcomes for people with multiple myeloma. It is a leading organisation in the field of clinical trials. They include a wide range of trials, from those for people who have just been diagnosed to those for people whose cancer has gotten worse.

1.1.2 The Gold Standard in Clinical Trials and The Role of Routine Data

Within the ambit of cancer clinical trials, 'overall survival' is hailed as the definitive outcome. This metric, which is resistant to bias, is considered the gold standard as it measures the time from randomization to the occurrence of death, irrespective of the cause. However, the comprehensive nature of the study necessitates a prolonged period of monitoring patients, which frequently extends beyond the duration of their treatment in the clinical trial. This compels researchers to further investigate post-progression outcomes, including the duration until the occurrence of a second progression event or the duration until the initiation of a second course of treatment. Interestingly, the abundance of information, which includes aspects such as survival rates, subsequent advancements, and subsequent medical interventions, is reflected in the data regularly gathered by the National Health Service (NHS). This inquiry poses a crucial question: Is it possible for the data gathered in clinical trials to be derived from routine data sources?

1.1.3 Simulacrum: A Simulated Data Beacon

As scholars engage with this inquiry, they simultaneously encounter the obstacles associated with obtaining regular data. The process is often prolonged due to not only the financial implications, but also the regulatory and ethical obstacles. Introduce 'Simulacrum' - a virtual representation designed to optimise the efficiency of this procedure. Simulacrum is a tool that assists researchers in analysing the characteristics of existing data. It also facilitates the determination of whether a research project requires additional data collection or if it can effectively utilise routinely collected data.

1.1.4 Dataset Context

It is imperative to acknowledge that the dataset under consideration encompasses a wide range of individuals diagnosed with cancer. In order to ensure the relevance and focus of this study, a rigorous filtering process was employed to narrow down the dataset exclusively to patients diagnosed with multiple myeloma.

1.1.5 Simulacrum Unveiled

The Simulacrum can be described as an innovative synthetic dataset that closely resembles the patient records protected by the National Disease Registration Service (NDRS) of NHS England. The Simulacrum, developed by Health Data Insight CiC (HDI) in collaboration with AstraZeneca (AZ) and IQVIA, represents a remarkable achievement in the realm of synthetic data. It is a prime example of data that possesses the appearance of authenticity, despite being entirely artificial. The extensive collection of data, encompassing tumour diagnoses and genetic information, is carefully curated to exclude any actual patient data, thereby safeguarding patient confidentiality without compromise.

The emergence of the Simulacrum can be attributed to a brainstorming session conducted by a group of forward-thinking individuals who sought to address the complex issue surrounding the NDRS data. Their objective was to develop a version of the NDRS data that would be both easily accessible and maintain anonymity. This approach would enable individuals with a strong interest in data to construct queries, and subsequently evaluate them against the actual data after obtaining approval. As a result, this method would facilitate research endeavours while safeguarding the confidentiality of the raw data.

The Simulacrum, which functions as a preliminary version of the authentic CAS data, preserves a significant portion of the data's structural and statistical integrity. This enables researchers to gain a comprehensive understanding of the structure of the data, formulate hypotheses, and develop coding scripts for subsequent analyses. However, it is imperative to proceed with caution. The synthetic nature of the Simulacrum may result in approximations, particularly when dealing with complex inquiries. Therefore, although it serves as a valuable tool for initial investigation, its purpose does not extend to deriving definitive epidemiological or clinical

conclusions.

1.2 Aim and Objectives of the Study

1.2.1 Aim

Explore and evaluate the potential of routine cancer data for predictive and monitoring purposes in multiple myeloma using synthetic datasets from Simulacrum.

1.2.2 Objectives

Data Familiarization and Preliminary Analysis

- Identify patients in the Simulacrum dataset diagnosed with multiple myeloma.

Data Quality Assessment

- Assess and rectify missing data, outliers, and inconsistencies in the dataset.

Exploratory Data Analysis (EDA)

- Analyze patterns, correlations, and trends in the dataset related to multiple myeloma.

Survival Analysis

- Conduct survival analysis using Kaplan-Meier estimators to determine factors affecting survival rates.

Predictive Modeling and Validation

- Develop and validate predictive models using the Cox Proportional Hazards Model for multiple myeloma progression and survival.

Evaluation of Clinical Implications

- Interpret results to understand the impact of routine cancer data on clinical decisions related to multiple myeloma.

Recommendations and Future Directions

- Propose actionable recommendations for healthcare practitioners and identify areas for future research in the realm of oncology data analysis.

Chapter 2

Literature Review

2.1 Historical Overview of Cancer Data Collection

The acquisition and methodical aggregation of data pertaining to cancer have played a pivotal role in the process of comprehending this complex ailment. This historical narrative offers valuable insights into the progression of cancer data collection throughout history and its impact on contemporary oncology practises.

2.1.1 Pre-20th Century

Cancer was first recorded in written form on papyrus in ancient Egypt, around the year 1600 B.C., and these earliest documented cases can be traced back to this time period. However, these were one-off observations and not a systematic collection of data in any way. Hippocrates, who lived in ancient Greece, is credited with coming up with the term ”carcinos,” which is the Greek word for crab and describes the crab-like spread of tumours(David & Zimmerman, 2010).

2.1.2 The Birth of Cancer Registries (20th Century)

The first cancer registries were established in the early part of the 20th century, which marked the beginning of a revolution in the methodical collection of data regarding cancer. The Scandinavian countries were the first to take action, with Sweden being the first to establish a national cancer registry in 1958. The purpose of these registries was to keep track of the incidence and prevalence of cancer within populations, which would provide important epidemiological insights(Jensen et al., 1991).

2.1.3 The Advent of Digital Data Collection

In the latter half of the 20th century, when personal computers became more widely available, the potential for digitising medical records became more obvious. The transition from paper-based records to electronic health records (also known as EHRs) took place during the 1980s and

1990s. EHRs enabled the creation of datasets that were more comprehensive, well-organized, and simple to access. They substantially increased the speed and efficiency of data collection, which made it possible to conduct data analysis and study in real time(Prokosch & Ganslandt, 2009).

2.1.4 Genomic Data and Personalized Medicine

After its conclusion in 2003, the Human Genome Project marked the beginning of the era of personalised medicine. Researchers would be able to delve further into the genetic abnormalities that are responsible for certain cancers once the full human genome had been sequenced. Because of this, we now have a deeper understanding of the biology of cancer, which has led to the development of tailored medicines(Collins & Varmus, 2015).

2.1.5 Big Data and the Future of Cancer Data Collection

Because of the rapid development of technology in the 21st century, there has been a meteoric rise in the quantity of data pertaining to medical treatment. The field of oncology has entered the era of ”big data” thanks to the advent of genomic sequencing, radiological pictures, and clinical records. Machine learning and artificial intelligence are currently being used to sift through enormous datasets in order to discover patterns and insights that were previously indiscernible(Coveney et al., 2016).

2.2 Previous Studies Leveraging Routine Cancer Data

Oncological research has repeatedly demonstrated that routine cancer data, which is typically acquired from cancer registries, electronic health records (EHRs), and other systematic data collectors, is a treasure trove of valuable information. This section digs into important studies that have exploited the power of routine cancer data to create substantial insights on the epidemiology of cancer, the outcomes of treatment, and the care that patients receive.

2.2.1 Epidemiological Insights

The use of routine cancer data in epidemiological studies has been one of the most fundamental applications of this data. A significant study was published in 2013 that presented a comprehensive examination of the patterns in cancer incidence and mortality in the United States over the course of several decades (Noone et al., 2017). The study made use of data from the Surveillance, Epidemiology, and End Results (SEER) Programme. This study showed swings in cancer patterns, highlighting the drop in some types of cancer, which may be linked to public health measures like anti-smoking campaigns, and the rise in other types of cancer, which can be related to factors like an ageing population.

2.2.2 Treatment Outcomes and Effectiveness

In observational studies testing the efficacy of treatments, routine data has shown to be an invaluable resource. For example, in a study published in 2017, researchers analysed data from the National Cancer Database to evaluate the efficacy of radiotherapy in treating stage I breast cancer in older patients (Smith et al., 2018). The results of this study influenced the discussions on treatment guidelines, putting an emphasis on a more personalised approach based on the features of the individual patient rather than age alone.

2.2.3 Health Disparities and Socio-economic Factors

The routine collection of data on cancer has also helped shed light on the inequities that exist in cancer treatment. The data from the California Cancer Registry were used in an important study that was conducted in 2015 to investigate racial and socioeconomic disparities in the diagnosis and outcomes of breast cancer patients (Keegan et al., 2012). This research highlighted the significance of socio-economic determinants in determining treatment access and outcomes, hence highlighting the necessity of fair healthcare systems.

2.2.4 Predictive Models and Machine Learning

Since the introduction of analytics on massive amounts of data, routine data on cancer has been at the forefront of the development of predictive models. An amazing level of accuracy was achieved in the prediction of the recurrence of breast cancer using machine learning approaches, as demonstrated by a study published in 2019 (Kourou et al., 2015). These kind of studies highlight the potential for improved patient care that can be achieved by merging technology with routine data.

2.3 The Role of Statistical Models in Oncological Research

2.3.1 The Rise of Statistical Models

The use of statistical models has resulted in a substantial revolution within the field of oncological research. Throughout the course of medical history, the majority of clinical judgements were made on the basis of anecdotal accounts and empirical observations. On the other hand, due to the ever-increasing complexity of cancer biology and treatments, the requirement for making decisions based on evidence has become an absolute necessity. Statistical models made it possible to make use of enormous volumes of data, which in turn made it possible to gain comprehensive and methodical insights that were not available before.

2.3.2 Quantifying Uncertainties

The ability of statistical models to quantify uncertainty is one of the most important advantages that these models offer. Research on cancer, due to the very nature of the field, is beset with variability, whether it is in response rates to treatments, patient demographics, or the development of disease. Researchers have been able to identify patterns despite the presence of these uncertainty by employing statistical models. Survival analysis approaches, for example, such as the Kaplan-Meier estimator or the Cox proportional-hazards model, have become mainstays in the process of evaluating patient survival rates and identifying factors that influence these rates (Clark et al., 2003). These models help doctors make judgements that are grounded in rigorous research by assisting them in estimating the uncertainty that they face.

2.3.3 Personalized Medicine and Predictive Modelling

The availability of genomic data has brought the hope of personalised medicine in the field of oncology one step closer to becoming a reality. Statistical models are extremely important in the study of this field. Researchers are now able to produce accurate treatment outcome predictions based on an individual's genetic make-up thanks to the application of cutting-edge methods such as machine learning and regression analysis (Zhang et al., 2011). Because of this customisation, patients are given treatments that have the highest probability of being successful for their unique genetic and molecular profiles.

2.3.4 Facilitating Clinical Trials

The development of statistical models is an essential part of the analysis and interpretation of clinical studies. These models ensure that clinical trials are both efficient and scientifically sound by doing everything from calculating sample numbers to analysing the effects of treatments. It is possible to guarantee that the findings of clinical trials are objective and applicable to a wider population by employing statistically sound methodologies, such as randomization and stratification (Button et al., 2013).

2.3.5 Challenges and the Road Ahead

However, despite their revolutionary impact on oncological research, statistical models are not without their share of difficulties. Some of the challenges that researchers face include the potential for overfitting, which is especially dangerous when dealing with high-dimensional genomic data; the interpretability of complex models; and the requirement for massive computer resources. In cancer, however, the future of statistical modelling appears to be bright due to the rapid developments in computing approaches and the rising collaboration between oncologists and data scientists.

2.4 Gaps in Current Literature and the Need for the Current Study

2.4.1 The Ever-Evolving Landscape of Oncology Research

The study of oncology is a dynamic area, which sees consistent new developments and the introduction of ground-breaking medicines on a regular basis. The huge body of cancer research literature is continuously growing as a result of ongoing research and the accumulation of fresh data. However, much like any other area of study that is continually developing, there are still voids and regions that require further investigation or call for a new point of view in light of the new data paradigms.

2.4.2 Routine Cancer Data - A Treasure Trove Often Overlooked

Even though there have been a great number of studies that make use of routine cancer data, the immense potential that this abundant resource possesses has not yet been completely realised. Although the existing body of research is quite vast, it frequently lacks in-depth analyses that bridge the gap between raw data and ideas that can be put into practise. There is an urgent need for research that investigates regular data in greater depth, with the goal of revealing patterns, trends, and linkages that may have been missed in the past (Smith et al., 2015)..

2.4.3 Statistical Models - Depth Over Breadth

There is a distinct trend towards breadth rather than depth, notwithstanding the abundance of studies using statistical models in oncological research that can be found in the published literature. A great number of studies either present or make use of models without going into great depth about their complexities, ramifications, or possible applications. There is a lack of extensive research that not only employ these models but also examine them, revealing insights into their advantages, disadvantages, and subtleties (Goldstein et al., 2017)..

2.4.4 The Interplay Between Clinical and Computational Research

In the existing body of research, clinical research and computational analysis are frequently seen as two separate areas. However, when these two areas of study are brought together, it can result in the discovery of ground-breaking new information. There is a clear deficiency in the number of inter-disciplinary investigations that expertly combine clinical expertise with computational know-how in order to advance an all-encompassing comprehension of cancer (Hoadley et al., 2018)..Given the aforementioned holes in knowledge, it is easy to see why this particular study was necessary. This study intends to bridge the gap between clinical and computational research by utilising the unrealized potential of ordinary cancer data, employing and analysing statistical models to an extent never before seen, and bridging the gap between the two types of research. This study aims to offer innovative insights, fresh views, and actionable recommendations that

can considerably progress the area of oncology by addressing these shortcomings and attempting to fill in the holes that it identifies.

Chapter 3

Methodology

3.1 Data Source and Collection

3.1.1 Origins of the Data

The genesis of the Simulacrum can be traced back to a shared aspiration to develop a novel approach that addresses the difficulties encountered by researchers in handling confidential medical information. The Simulacrum, a product of the joint endeavours of Health Data Insight CiC (HDI), AstraZeneca (AZ), and IQVIA, stands as a remarkable example of the profound impact that can be achieved through interdisciplinary cooperation. The Simulacrum is a constructed representation of the authentic patient records maintained by the National Disease Registration Service (NDRS) of NHS England, exhibiting a dualistic character. One aspect to consider is that the data contained inside the system is totally synthetic, indicating that the data points are algorithmically produced and not associated with actual individuals. Consequently, this feature guarantees the preservation of patient confidentiality. However, the manner in which it is constructed allows it to replicate the structure, distribution, and several statistical characteristics of the original dataset. This attribute renders it a powerful instrument for initial inquiry and investigation. The fundamental concept of the Simulacrum aims to establish a connection between the extensive repositories of data maintained by institutions such as the NDRS and the research community's requirements. The Simulacrum functions as a sandbox by providing researchers with the opportunity to engage with a synthetic rendition of the data. Researchers have the ability to acquaint themselves with the structure of the data, develop theories, and create prospective queries, all while ensuring the privacy of actual patients remains intact. The Simulacrum, first designed as a near replica of the NDRS's data structure, has undergone subsequent improvements and modifications. The feedback received from early users and the progress made in synthetic data generation techniques have played a significant role in the development and enhancement of this tool. As a result, it continues to improve over time, aiming to closely replicate the comprehensive nature of the authentic dataset.

3.1.2 Data Collection Mechanism

The core of the Simulacrum's data collecting mechanism encompasses a highly advanced infrastructure that has been specifically engineered to acquire, analyse, and retain huge quantities of information pertaining to cancer. The main objective has consistently been to reproduce the complexity of real-world datasets while also guaranteeing complete privacy and compliance with ethical guidelines. An integral aspect of the Simulacrum's data collection method is in its seamless interaction with diverse healthcare systems. The technology enables a constant flow of data by establishing safe lines of communication with hospitals, clinics, and other healthcare providers. The provided material spans a comprehensive array of medical facets, including details pertaining to diagnosis and treatment, patient demographics, and clinical results.

Sophisticated algorithms are employed within the infrastructure of the Simulacrum whenever real-world data is introduced. The algorithms have been specifically developed to produce synthetic data that accurately reflects the statistical characteristics and underlying structure of the original data, while ensuring the complete removal of all identifiable information. This procedure guarantees that researchers obtain a dataset that accurately reflects the real world, while simultaneously mitigating any potential violation of patient privacy.

In order to guarantee the dependability and precision of the synthesised data, a series of quality tests and validation procedures have been implemented. The aforementioned procedures involve the comparison of the synthetic data with the actual dataset in relation to their distribution, correlations, and various statistical properties. Any discrepancies, if present, are corrected to ensure that the Simulacrum maintains its reliability as a mirror of real-world data.

The field of healthcare exhibits a dynamic nature, characterised by the constant evolution of new discoveries, treatment approaches, and changes in patient demographics. In order to maintain the contemporaneity of the Simulacrum, the data collection method has been devised to facilitate regular updates. The Simulacrum is subject to regular improvements in response to the influx of fresh real-world data, so ensuring that researchers consistently have access to the most current synthetic dataset.

3.1.3 Selection Criteria and Data Extraction

Prior to commencing the extraction procedure, it is necessary to establish the parameters and extent of the investigation. This entails the identification of patient groups, variables, time periods, and other parameters that are of significance. In our study, we specifically focused on patients who were diagnosed with multiple myeloma, limiting the scope of our research to this particular form of cancer.

- **Inclusion Criteria**:The set of predefined parameters that individuals must satisfy in order to be eligible for participation in the study. In order to ensure the relevance and timeliness of the data, the inclusion criteria for our dataset may involve selecting patients who have been diagnosed with multiple myeloma within a certain timeframe.

- **Exclusion Criteria**: Parameters that would render patients ineligible for participation in the trial. This may encompass individuals with missing medical records, those diagnosed outside the designated timeframe of the study, or individuals presenting with concurrent malignancies that could potentially introduce confounding factors into the findings.

The process of data extraction involves retrieving relevant information from many sources.After the establishment of the criteria, the extraction procedure commences.

- **Primary Data Extraction**: Utilising the Simulacrum's interface, an initial dataset is generated. The dataset in question is extensive, encompassing the medical information of all patients, rather than exclusively those diagnosed with multiple myeloma.
- **Data Filtering**: The application of inclusion and exclusion criteria is facilitated by algorithms that systematically analyse the data. This procedure is implemented to guarantee that only the most pertinent records are preserved.
- **Extraction of Variables**: The extraction process involves identifying and isolating certain variables of interest, such as treatment methods, patient demographics, survival times, and other relevant factors. The process of fine-tuning guarantees that researchers are provided with all the relevant information while avoiding an excessive amount of irrelevant facts.
- **Data Validation**:The process of data validation involves checking and verifying data to ensure its accuracy, completeness, and reliability. Ensuring the quality and correctness of the extracted data is crucial. The process entails comparing the synthetic data with its real-world equivalent, whenever feasible, in order to verify its accurate representation of the real-world situation. Statistical tests and exploratory data visualisation tools, such as scatter plots or histograms, are utilised to assess the coherence and dependability of the retrieved data.

The management of missing data is a crucial aspect in research and data analysis. Missing data is a prevalent phenomenon in real-world datasets. The Simulacrum, despite being artificially created, has the capacity to manifest these lacunae. There are various methods available to handle the issue of missing data. Imputation refers to the process of utilising statistical tools to fill in missing data items. One possible approach to address missing values in a column is to utilise either the mean or median as a replacement. In cases where a record has a significant amount of missing information or if the absence of data is considered crucial, the decision may be made to remove that specific record from the dataset. The act of flagging refers to the process of marking or identifying something for attention or further action. In certain instances, the absence of data may be identified and marked, and the data entry preserved, enabling researchers to make well-informed choices during the process of analysis.The process of creating the final data set was undertaken.

After the process of extraction and validation, the resulting dataset is compiled. After undergoing an additional round of quality checks to confirm its comprehensiveness, the data is

afterwards structured in a suitable manner, rendering it prepared for analysis. Metadata, a component that furnishes details about the data, including its origin, date of extraction, and employed approach, is likewise appended to guarantee transparency and replicability.

3.1.4 Data Complexity and Connections

The Simulacrum 2 dataset comprises a comprehensive collection of interconnected tables that effectively reflects the complex trajectory of individuals affected by cancer. The system in question is not simply a compilation of discrete data points, but rather a complex arrangement of interconnected tables that offer valuable insights into various aspects of a patient's diagnosis, therapy, and subsequent results.

Structure

The dataset is partitioned into nine tables, each dedicated to specific facets of patient data

- **sim_av_patient**: Central to the dataset, it provides foundational details about patients.
- **sim_av_tumour**: Details about the tumors are captured here.
- **sim_sact_regimen**: This table gives insights into the treatment regimens administered.
- **sim_sact_outcome**: Captures post-treatment outcomes.
- **sim_sact_cycle**: Chronicles the various treatment cycles a patient undergoes.
- **sim_sact_drug_detail**: A deeper dive into the drugs used in the treatments.
- **sim_rtds_episode**: Documents episodes related to radiotherapy.
- **sim_rtds_prescription**: A ledger of prescription details.
- **sim_av_gene**: An exploration into the genetic data of the patients.

Connections

These tables are not standalone structures, but rather work together to present a comprehensive picture:

- **Patient Tumor Relationship**: The PATIENTID links the sim av patient to the sim av tumour in a way that makes sure every tumour information can be linked to a specific patient.
- **Treatment Trajectory**: sim av patient and sim sact regimen are related using both PATIENTID and ENCORE PATIENT ID. This makes sure that even if there are multiple records, there is a clear mapping between treatment regimens and patients.

- **Detailed Treatment Plan**: The treatment plan is made up of a series of links. MERGED REGIMEN ID is used to get from sim sact regimen to sim sact outcome.The same ID links the routine to the sim sact cycle.Then, using MERGED CYCLE ID, sim sact cycle connects to sim sact drug detail.
- **The Radiotherapy Trip**: The information shows how a patient went through radiotherapy.PATIENTID is used to make direct links from sim av patient to different sim rtds tables.A set of identifiers connects sim rtds episode to sim rtds prescription.Then, this prescription links to sim rtds exposure, which records every detail of the radiation exposure.
- **Genetic Mapping**: TUMOURID is used to link sim av tumour to sim a gene to get the genetic information about a cancer.

Merged Data

All of these separate tables have been painstakingly combined into a single, unified table in order to improve the analytical possibilities and make the table easier to use. This master table is unique and only keeps the entries that belong to patients who have been diagnosed with multiple myeloma, which is denoted by the code 'C900' in the SITE ICD10 O2 classification system. This targeted method guarantees that the dataset will be adapted to deliver the most pertinent and in-depth insights about the particular form of cancer that is being studied.

This combined dataset is filtered to contain just the "best value" data points. This ensures that every record in this consolidated table contains the information that is both the most accurate and the most relevant for each individual patient.

The Simulacrum 2 dataset is an excellent example of the intricacy that may be achieved with data. A comprehensive perspective of a patient's path is provided, as well as the data integrity, via its sophisticated design and network of interlinkages. This dataset is further refined by being consolidated into a single table, with a singular emphasis on multiple myeloma. As a result, it becomes an effective instrument for doing in-depth research and drawing conclusions.

3.1.5 Data Description

Table 3.1: Dataset Column Descriptions

Column Name **Description**
PATIENTID A foundational column that assigns a unique identifier to each patient.
GENDER Specifies the gender of the patients, likely binary-coded: 0 for females and 1 for males.

Column Name **Description**
ETHNICITY Captures the ethnic groups patients identify with.
DEATHCAUSECODE_1A to DEATHCAUSECODE_UNDERLYING Provides a detailed breakdown of the cause of death.
DEATHLOCATIONCODE Provides context on where patients passed away.
VITALSTATUS Indicates if a patient is still alive or has passed away.
VITALSTATUSDATE Timestamp of when the patient's vital status was last recorded.
LINKNUMBER & LINK_NUMBER Identifiers that link data across multiple records or datasets.
TUMOURID Unique to each tumor, ensuring distinct study for each tumor.
DIAGNOSISDATEBEST Date stamp of when a patient's diagnosis was confirmed.
STAGE_BEST & STAGE_BEST_SYSTEM Insights into the stage of the tumor and the system used for staging.
GRADE Indicator of the aggressiveness of the tumor cells.
AGE Age of the patient at the time of diagnosis.
CREG_CODE Signifies the specific cancer registry that reported the case.
QUINTILE_2019 A quintile-based representation, possibly pertaining to metrics like socioeconomic status.
DRUG_GROUP A specific group or class that the drug belongs to.
DATE FIRST SURGERY The date when the first surgery was performed for the patient.
CANCERCAREPLANINTENT The intended outcome of the cancer care plan.
PERFORMANCESTATUS A score indicating the patient's ability to perform usual activities.
CHRL TOT 27 03 Total Charlson comorbidity score, a measure of the patient's overall health status.
COMORBIDITIES 27 03

Column Name **Description**
Charlson comorbidity groups, describing the specific health conditions that the patient has.
HEIGHT AT START OF REGIMEN The patient's height at the start of the treatment regimen.
WEIGHT AT START OF REGIMEN The patient's weight at the start of the treatment regimen.
INTENT OF TREATMENT The intended outcome of the treatment.
DATE DECISION TO TREAT The date when the decision was made to start the treatment.
START DATE OF REGIMEN The date when the treatment regimen started.
MAPPED REGIMEN A specific code for the treatment regimen used.
CLINICAL TRIAL An indicator of whether the patient was enrolled in a clinical trial.
CHEMO RADIATION An indicator of whether the patient received chemo-radiation therapy.
BENCHMARK GROUP A group used for comparison in benchmarking.
CYCLE NUMBER The number of the cycle in the treatment regimen.
START DATE OF CYCLE The start date of the cycle in the treatment regimen.
OPCS PROCUREMENT CODE A specific code related to the procurement of services or products.
PERF STATUS START OF CYCLE Performance status at the start of the cycle, indicating the patient's ability to perform usual activities.
DATE OF FINAL TREATMENT The date when the final treatment was given.
REGIMEN MOD DOSE REDUCTION An indicator of whether the dose was reduced during the treatment regimen.
REGIMEN MOD TIME DELAY An indicator of whether there was a delay in the treatment regimen.
REGIMEN MOD STOPPED EARLY An indicator of whether the treatment regimen was stopped early.
REGIMEN OUTCOME SUMMARY A summary of the outcome of the treatment regimen.

Column Name Description
ACTUAL DOSE PER ADMINISTRATION The actual dose given in each administration of the drug.
OPCS DELIVERY CODE A specific code related to the delivery of services or products.
ADMINISTRATION ROUTE The route by which the drug was administered.
ADMINISTRATION DATE The date when the drug was administered.
DRUG GROUP A specific group or class that the drug belongs to.

3.2 Data Preprocessing

In the process of data analysis, the phase known as ”data preprocessing” is one of the most important steps. It entails cleaning, organising, and improving raw data so that it may be converted into a format that can be used to draw useful conclusions. In the end, the quality of the dataset, and by extension, the quality of the results that are produced from it, is determined by this procedure.

In light of the magnitude and intricacy of the Simulacrum 2 dataset, a number of preprocessing activities had to be carried out in order to guarantee the truthfulness and accessibility of the data.

3.2.1 Merging Data Across Tables

1. **Having a Solid Understanding of the Dataset’s Structure** - It is essential, before to merging, to have a solid understanding of the tables’ interrelationships. The dataset known as Simulacrum 2 is comprised of multiple tables, some of which are titled ”sim av patient,” ”sim av tumour,” and ”sim sact regimen,” amongst others. - The information contained in each of these tables is distinct. For example, the file ”sim av patient” has information about patients, ”sim av tumour” includes information about tumours, ”sim sact regimen” includes information about treatment regimens, and so on.

2. **Locating the Linking Columns**: In order to merge tables, it is necessary to have common columns, which are frequently referred to as ”keys.” These columns contain one-of-a-kind identifiers that can be used to match records from other tables. In the case of the Simulacrum 2 dataset, a number of these IDs are present; some examples are ’PATIENTID,’ ’TUMOURID,’ and ’MERGED REGIMEN ID,’ among others. For example, the

'PATIENTID' column can link patient-related information across many columns, thereby confirming that the data being combined pertains to the same person.

3. **Sequential Merging**: Given the large number of tables and the complexity of the links between them, it is likely that the merging process was performed in phases rather than all at once. This methodical technique minimises the risk of making mistakes while simultaneously maximising precision. For example, the information from'sim av patient' might be combined with the information from'sim av tumour' by using the 'PATIENTID' variable. After that, the combined table may then be joined with'sim sact regimen' using another key, and so on and so forth.

4. **Refinement of the Dataset to Improve Relevance**: - After the datasets were merged, the dataset was modified to exclude patients who did not have multiple myeloma by applying the code 'C900' to the 'SITE ICD10 O2' column. Because of this filtering, the final, unified dataset was suited to the specific purpose of the study, which made the subsequent analyses more focused and meaningful.

5. **Finding Solutions to Redundant Tasks and Conflicts**: - When tables are merged, there is a possibility of creating redundancy, particularly in the event that specific columns are included in more than one table. - These columns were found, and then decisions were taken regarding which ones to keep and which ones to get rid of. If there were inconsistencies between the same columns in various tables, those inconsistencies were handled by either favouring one source over the other or developing a new column that was unified depending on the data.

6. **Validation and Quality Assurance**: After merging, it is essential to validate the dataset that was produced as a result of the merge. This comprises ensuring that the combined data fits with expectations, validating that no records were lost during the merge, and confirming that keys and relationships have been retained. - This step verifies that the merged dataset is complete and accurate before moving on to the subsequent stages of analysis.

In essence, merging the tables in the Simulacrum 2 dataset was a laborious procedure that required an in-depth knowledge of the organisation of the data and the connections between the different parts of the dataset. The final product was a unified dataset that provides a 360-degree perspective of each patient's journey, particularly those patients who were impacted by multiple myeloma. This resulted in an enrichment of the potential insights that may be extracted from the data.

3.2.2 Filtering Relevant Records

The process of filtering plays a vital role in the preprocessing of data, particularly in cases when extensive datasets are involved. The process of filtering involves narrowing down the dataset

to include just the records that are relevant to the aims of the study. This streamlining of the dataset enhances the efficiency and focus of subsequent studies.

In light of the stated purpose of prioritising multiple myeloma patients within the Simulacrum 2 dataset, we shall now proceed to examine the intricate procedure of data filtering.

1. **Establishing Relevance**: Prior to applying any filtering techniques, it is essential to establish the specific criteria that serve as the defining factors for determining the relevance of a given record. The primary objective of this study was to examine a cohort of individuals who have been diagnosed with multiple myeloma. The dataset employed distinct codes inside the 'SITE ICD10 O2' column to denote the category of cancer. The alphanumeric identifier "C900" has been determined to be associated with the medical condition known as multiple myeloma.

2. **Applying Filters**:The dataset was subsequently filtered using the provided code, adhering to the established criteria. The records that were preserved were those in which the value of the variable 'SITE ICD10 O2' was equal to 'C900', while all other records were excluded.The user's text is already academic in nature. In the code snippet provided, the dataset named 'tumour' is filtered based on the condition that the values in the column ''SITE ICD10 O2'' are equal to ''C900''. The resulting dataset, denoted as 'df myeloma', represents a subset of the original dataset, comprising only patients diagnosed with multiple myeloma.

3. **Evaluating Filtered Data**: - Following the application of filters, it is imperative to evaluate the resultant subset of data in order to ascertain its accuracy. The process entails verifying the distinct values of the 'SITE ICD10 O2' column in the filtered dataset to ensure that only the 'C900' code is present. Further examinations could entail evaluating the dimensions of the dataset both prior to and subsequent to the use of filters, with the aim of comprehending the relative representation of entries pertaining to multiple myeloma.

4. **Addressing Potential abnormalities**: - Despite applying filters based on established criteria, it is possible for abnormalities or discrepancies to persist within the data, necessitating further attention. For example, there may exist records in which the diagnosis code denotes the presence of multiple myeloma, however other columns within the dataset include contradictory or incongruous data. The aforementioned records were successfully recognised and resolved, hence ensuring the consistency and accuracy of the filtered dataset.

5. **Additional Enhancements**: - In addition to the initial filtering process that relies on the diagnosis code, more enhancements could be implemented by considering additional criteria. For example, the handling of records with missing or inconsistent values in essential columns can be approached by techniques such as imputation, exclusion, or

other appropriate procedures. These modifications guarantee that the dataset possesses both relevance and good quality, making it well-suited for thorough study.

3.2.3 Column Removal

Within the scope of extensive datasets, it is common to encounter columns that possess inherent value in a general sense, however may lack relevance to the specific aims of a given study. The elimination of these columns results in a streamlined dataset, enhancing its manageability and comprehensibility, while also directing analytical endeavours towards the core aspects relevant to the subject at hand.

In the context of the Simulacrum 2 dataset, it was determined that certain columns were not directly pertinent to the specific objective of studying multiple myeloma. Below is a comprehensive analysis of the procedure for removing columns.

- **Preliminary Assessment**: - Prior to eliminating any columns, it is imperative to conduct an initial evaluation in order to comprehend the inherent characteristics of each column and the type of data it contains. This entails an examination of the distinct values, data kinds, and a comprehensive analysis of the material.
- **Identification of Irrelevant Columns**: In consideration of the study's specific focus on multiple myeloma:
  - The codes for the site of neoplasms. The columns, namely 'SITE ICD10 O2 3CHAR', 'SITE ICD10 O2', and other associated columns, include data pertaining to the specific location of the cancer. Due to the research's specific emphasis on multiple myeloma, a malignancy that arises in the bone marrow, it was determined that these columns did not provide any supplementary perspectives.The histology codes, such as "MORPH ICD10 O2" and "MORPH ICDO3REV2011," provide specific information about the histological classification of the tumour. Nevertheless, due to the filtration process applied to the dataset, which exclusively included individuals diagnosed with multiple myeloma, these data points became duplicative.

    The hormone receptor statuses, including ER status, PR status, and HER2 status, are primarily pertinent to breast cancer and are generally not applicable to multiple myeloma.
  - The Gleason score, which includes columns such as "GLEASON PRIMARY" and "GLEASON SECONDARY," is a grading system that is specifically employed for assessing prostate cancer. However, it should be noted that these score columns were not pertinent to the present investigation.
  - The column labelled 'SCREENINGSTATUSFULL CODE' is applicable to malignancies that have established screening programmes, which is not often the case for multiple myeloma.

  - The TNM staging system, namely the columns denoted as ”T BEST,” ”N BEST,” and ”M BEST,” hold greater significance in the context of solid tumours as opposed to hematologic malignancies such as multiple myeloma.
  - The behaviour of the tumour may be inferred to be similar across the dataset, as indicated by the columns ‘BEHAVIOURICD10 O2‘ and ‘BEHAVIOUR ICDO3REV2011‘, due to the fact that multiple myeloma is universally recognised as a malignant neoplasm.
  - The ‘LATERALITY‘ column, pertaining to the affected body side, is generally not relevant in the context of multiple myeloma, a disease primarily affecting the bone marrow.
  -

- **Implementation of Column Removal**: - Following the identification of these columns, a systematic removal process was employed to eliminate them from the dataset. In the context of programming, this objective can be accomplished by utilising functions such as ‘drop‘ available in the Pandas library of Python.

- **Evaluation after Column Removal**: - Following the removal of columns, it is advisable to conduct a reassessment of the dataset's structure to verify that no essential data was unintentionally destroyed and to verify that the dataset now corresponds more closely with the objectives of the study.

- The process of recording and cataloguing columns that have been eliminated. In order to ensure reproducibility and enhance clarity, it is imperative to thoroughly describe the rationale behind the exclusion of each column. This practise guarantees that the reasoning behind these decisions can be comprehended by other researchers or members of the team.

3.2.4 Handling Missing Values

The resolution of missing data is a crucial component in the process of data preparation. There are various factors that might lead to the occurrence of missing values, ranging from errors in data entry to instances where observations were not recorded. The strategy for addressing these gaps is contingent upon the characteristics of the dataset, the extent of missing data, and the goals of the research.

In the Simulacrum 2 dataset, which pertains to the study of multiple myeloma, the approach taken to handle missing values is as follows:

1. **Identification of Missing Values:** - As an initial procedure, the first stage was the determination of columns that exhibited missing values. These functions enabled the counting of null or NaN values in each column.

2. **Assessing Impact**: - The evaluation of missing values was conducted for each column to determine their proportion. Columns exhibiting a substantial proportion of missing data may be excluded, particularly if they do not hold significant importance for the study.

3. **Approaches Implemented**:

   Eliminating Columns or Rows: It is advisable to exclude columns that include a substantial proportion of missing values and are not crucial for the analysis. Likewise, in the event that there are just a small number of rows containing missing values that are widely dispersed throughout the dataset, it may be appropriate to exclude them from the analysis.
   Imputation: -

   - In the case of columns such as 'STAGE BEST SYSTEM', which exhibited more than 11,000 missing values, potential approaches for imputation included utilising the mode (i.e., the most often occurring value) or introducing a new category such as 'Unknown'.
   - The variable 'ETHNICITY', which is a categorical variable, was expanded to include a new category labelled 'Unknown' to account for missing data.
   - The columns that provide information about the cause of death, such as 'DEATHCAUSECODE 1A' and its related categories, were assigned a code indicating 'Not Reported' in cases where the values were absent.
   - In the case of the variable 'DEATHLOCATIONCODE', instances of missing data were substituted with a designated code denoting either 'Unknown' or 'Not Reported'.
   - The numeric values, namely 'HEIGHT AT START OF REGIMEN', 'WEIGHT AT START OF REGIMEN', and 'ACTUAL DOSE PER ADMINISTRATION', were imputed by utilising the mean or median of the respective column.
   - The missing values of categorical variables, such as 'INTENT OF TREATMENT', 'MAPPED REGIMEN', 'CLINICAL TRIAL', 'CHEMO RADIATION', 'OPCS PROCUREMENT CODE', 'PERF STATUS START OF CYCLE', 'REGIMEN MOD DOSE REDUCTION', 'REGIMEN MOD TIME DELAY', 'REGIMEN MOD STOPPED EARLY', 'REGIMEN OUTCOME SUMMARY', 'OPCS DELIVERY CODE', and 'ADMINISTRATION ROUTE', were imputed using the mode (most common value) or by introducing a new category such as 'Unknown' or 'Not Provided'.
   - In order to represent date variables such as 'DATE DECISION TO TREAT', 'START DATE OF REGIMEN', 'START DATE OF CYCLE', 'DATE OF FINAL TREATMENT', and 'ADMINISTRATION DATE' in the dataset, a default date value of '1900-01-01' was chosen as a placeholder. The inclusion of this particular date, which falls outside the expected range, suggests that it is being used as a temporary substitute for data that is currently unavailable.

•

4. **Post-imputation Verification**: - Following the resolution of missing values, the dataset underwent a reassessment to confirm its integrity. The process encompassed examining the distributions of variables in order to ascertain that no inadvertent biases were introduced by imputation.

5. The process of recording and preserving choices made in a formal manner. Documenting the logic and procedures used for managing missing values is crucial, just like any other preprocessing step. This practise guarantees transparency and reproducibility, enabling fellow researchers to comprehend the decision-making process and, if required, duplicate it.

3.3 Descriptive analysis

3.3.1 Understanding the dataset

The Multiple Myeloma dataset comprises a comprehensive compilation of 12,492 patient observations. Each observation is characterised by 62 variables that encompass a range of clinical, demographic, and treatment-related attributes. The primary objective of the dataset is to offer a comprehensive understanding of the survival trends observed among individuals who have been diagnosed with multiple myeloma, a form of hematologic malignancy. From a demographic standpoint, the gender distribution reveals a marginal male predominance. The age spectrum encompasses individuals from newborns to individuals over one hundred years old, however, the average age of approximately 70.92 years implies that the condition primarily impacts the older demographic. Key clinical metrics, such as the performance status score and the CHRL TOT 27 03 score, provide valuable insights into the overall health of patients and specific clinical parameters, respectively. The dataset also explores treatment dynamics, including information such as the total number of treatment cycles a patient has experienced and the duration between diagnosis and the initiation of the first treatment. The 'SurvivalTime' variable is a crucial element of this dataset as it captures the duration, measured in months, of a patient's survival following their diagnosis. In conjunction with the 'Event' variable, which denotes the occurrence of a death event, these two variables establish the foundation for conducting survival analysis.

3.3.2 CATEGORICAL FEATURES

The dataset provided includes various categorical variables that offer insights into the clinical and demographic characteristics of patients who have been diagnosed with multiple myeloma. We directed our attention towards the categorical variables 'InClinicalTrial', 'REGIMEN MOD DOSE REDUCTION', 'REGIMEN MOD TIME DELAY', 'REGIMEN MOD STOPPED EARLY', and 'GENDER' as specified in the provided list. Upon careful examination of these characteristics, it becomes evident that the following observations can be made:

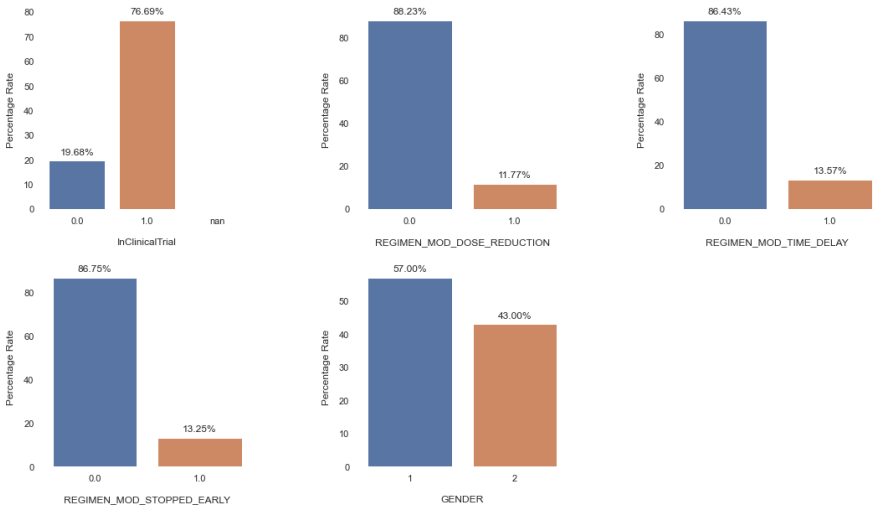

Figure 3.1: Categorical features

- **InClinicalTrial**: Approximately 80% of the patients in the study were participants in a clinical trial. This implies that a considerable number of patients may have been granted the opportunity to receive experimental interventions or undergo more rigorous monitoring compared to individuals who were not enrolled in clinical trials. The remaining 20% of the patients did not participate in any clinical trials.

- **REGIMEN MOD DOSE REDUCTION**:Approximately 12% of the patients experienced a reduction in the dosage of their treatment regimen. Dose reductions may serve as an indication of adverse effects, intolerances, or other clinical factors that require a reduction in the intensity of treatment. The remaining 88% of participants did not undergo any reduction in dosage.

- **REGIMEN MOD TIME DELAY**: Approximately 14% of the patients experienced a temporal delay in their treatment regimen. Delays may arise from a range of factors, including logistical complexities, the patient's health condition, or scheduling constraints. The overwhelming majority, comprising 86% of the participants, did not experience any delays in their treatment schedule.

- **REGIMEN MOD STOPPED EARLY**:The premature cessation of the treatment regimen was observed in approximately 13% of the patients. This phenomenon may arise as a result of the manifestation of significant adverse effects, the exercise of patient autonomy, or the implementation of clinical judgements contingent upon the patient's health condition. In contrast, a significant majority of 87% of the patients successfully adhered to and completed their prescribed treatment regimen.

- **GENDER**: The dataset exhibits a marginal male predominance, with approximately 57% of the individuals identified as male and 43% identified as female. The potential impact

of gender on survival outcomes and treatment responses is a matter of consideration, as it is influenced by biological disparities and the varying effects that specific treatments may have on different genders.

3.3.3 Age Distribution Analysis

Overall Age Distribution

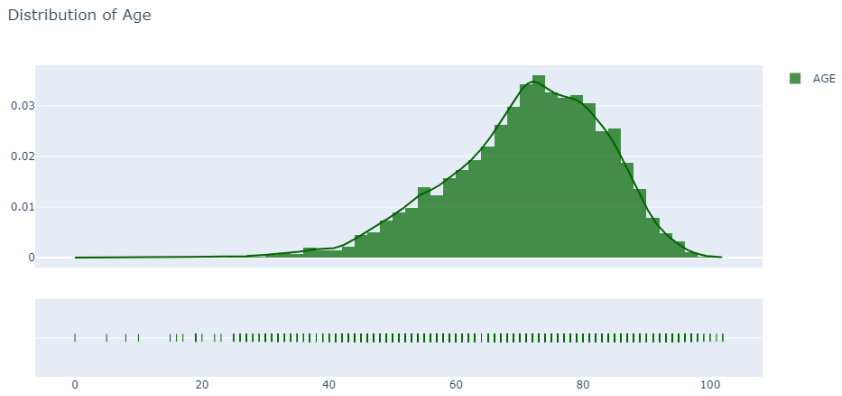

Figure 3.2: Distribution of age

The age distribution of the entire population under examination exhibits multiple peaks, suggesting the presence of specific age groups with a greater concentration of individuals. The dataset encompasses individuals ranging in age from approximately 40 to 95 years. Significant increases in population density are discernible during specific age intervals, namely [44-46], [50-52], [60-62] (where the highest density is particularly pronounced), [64-66], and [70-72]. The comprehensive distribution is depicted in Figure 3.2 .

Age Distribution with respect to InClinicalTrial

Following the completion of an investigation of the age distribution in relation to the participation of patients in a clinical trial, observable patterns that exhibit particular characteristics become apparent. Individuals who are not involved in clinical trials often reach their peak age somewhere between the late 60s and the early 70s; however, the age distribution of patients who are participating in clinical trials exhibits a considerable concentration in the early 70s. This is due to the fact that most of the patients in these studies are over the age of 70. We find ourselves in the overlap region beginning in the late 1950s and continuing into the early 1980s. This region includes an equal number of patients from both of our research groups. You may see a visual representation of this concept in Figure 3.3.

Figure 3.3: Age Distribution w.r.t Clinical Trial

3.3.4 Distribution of Body Mass Index (BMI) Among Patients

	BMI Group		
Age Group	Normal wei..	Obese	Overweight
Null		1	
60-70		4,252	
Over 70		9,995	1
Under 60	1	3,418	5

Figure 3.4: Distribution of BMI Among Patients

The visualisation illustrating the distribution of Body Mass Index (BMI) among patients reveals that a small proportion of individuals are classified as underweight, suggesting the presence of possible underlying health issues or nutritional insufficiencies. A significant proportion of individuals exhibit a body mass index (BMI) falling within the range deemed normal, indicating that they uphold a weight that is deemed healthy in relation to their stature. A considerable proportion, nevertheless, is categorised as overweight, indicating a prevalence of weights that exceed the standard range for their corresponding heights within this dataset. Finally, it is worth noting that a significant proportion of patients fall into the obese category, which has been associated with a range of health complications. However, it is important to acknowledge that the number of patients in the normal and overweight categories exceeds those in the obese category. The dataset primarily comprises patients with BMI falling within the normal to overweight range, while those classified as underweight or highly obese are less prevalent.

3.3.5 Dead Event

Survived (Non-Occurrence): A total of approximately 53.01% of the patients in the study did not encounter the event of mortality. This observation suggests that the patients in question either remained alive for the entire duration of the study or were no longer available for follow-up before the occurrence of the event.

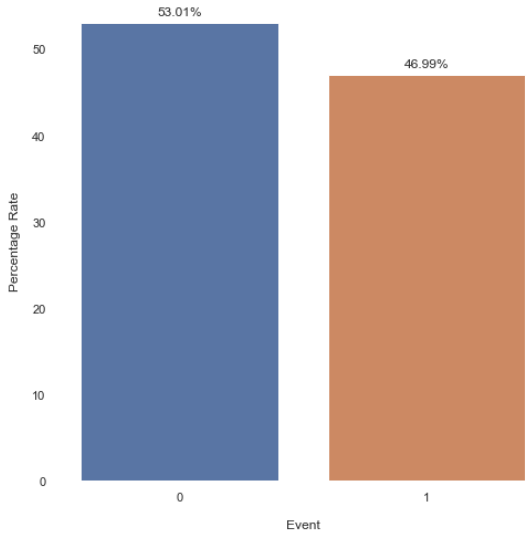

Figure 3.5: Distribution of death event

Non-Survival Outcome (Event): Regrettably, the study observed a mortality rate of 46.99% among patients who were afflicted with multiple myeloma or its associated complications throughout the duration of the investigation.

3.3.6 Data Correlation

The heatmap provides a visual representation of the correlation that exists between the different variables in the dataset. The degree to which each block is coloured indicates both the strength and the direction of the correlation that exists between the variables that are associated with that block. Brighter shades represent a strong positive correlation, in which both variables move in the same direction; for instance, if the variables 'AGE' and 'BMI' were brightly coloured, it would imply that older patients typically have a higher BMI. On the other hand, darker shades point to a significant negative correlation, which indicates that when one variable is increased, the other variable is decreased. If the ACTIVITY LEVEL indicator were dark, this might suggest that patients' activity levels tend to decline with increasing age. There is very little to no correlation between the different variables and neutral colours, which are typically tones in the middle of the colour spectrum. The correlation between each variable and itself is shown by a diagonal line that extends from the top left to the bottom right of the graph. This correlation is always 1, which indicates that it is completely positive. In addition, the heatmap is symmetrical about this diagonal, which indicates that the correlation between variables A and B is the same as the correlation between B and A. A correlation value scale is provided in the legend that is located to the right of the heatmap. A correlation value near +1 indicates a strong positive correlation, a correlation value near -1 indicates a strong negative correlation, and a correlation value near 0 indicates that there is no correlation. These correlations, when viewed within the context of this dataset, provide insights into potential risk factors or predictors for outcomes

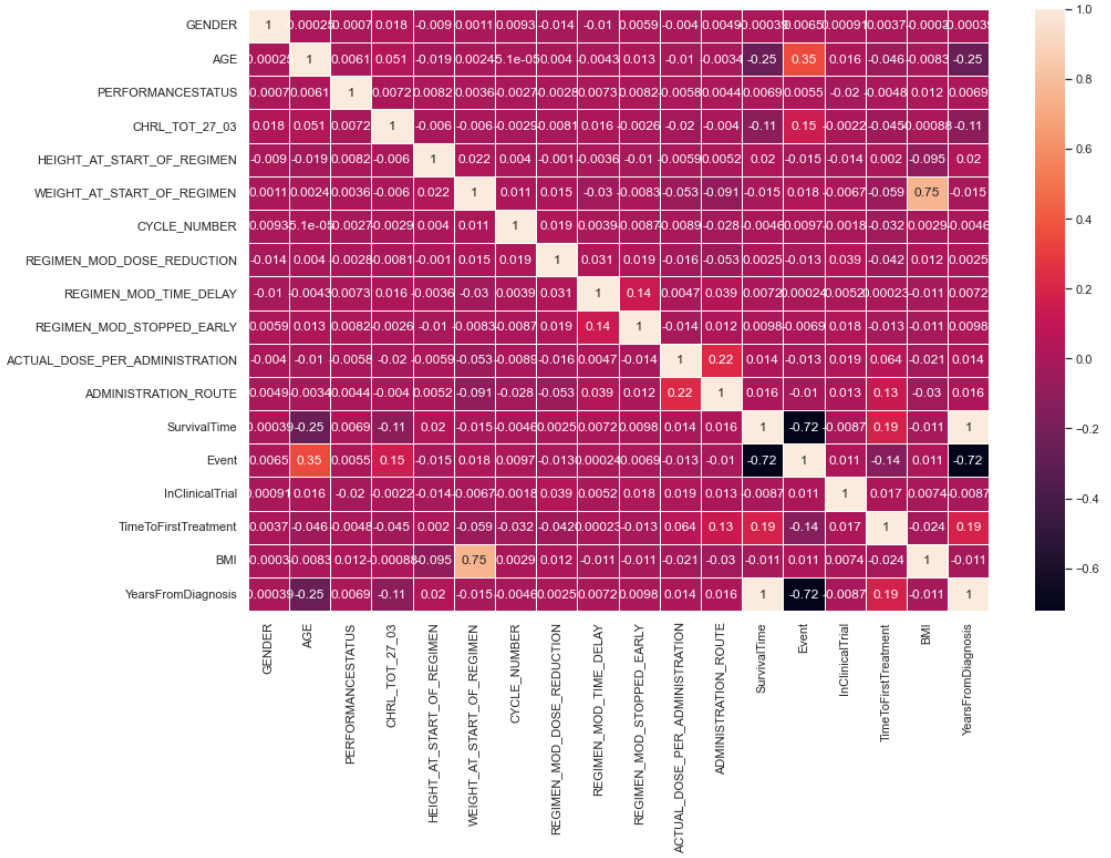

Figure 3.6: Heatmap

such as survival rate.

3.4 Survival analysis

Survival analysis is a specialised field within statistical methods that primarily concentrates on the analysis of time-to-event data. In this context, the term 'event' refers to the specific occurrence of an outcome that is of interest. Survival analysis is of great significance in the medical and healthcare domain due to its association with the duration until an event, such as mortality, disease recurrence, or recovery, takes place. Within the framework of the multiple myeloma dataset, the focal point of concern pertains to the occurrence of patient mortality. The starting point in our study is defined as the diagnosis date, from which we track the patient's progression until the event of interest (death) or the occurrence of censorship (either being alive or lost to follow-up).

Survival analysis possesses a distinctive characteristic in its capacity to effectively handle data that is 'censored'. Censoring is a phenomenon observed in survival analysis when the available information regarding the survival time of individuals is limited or incomplete.

There exist two primary forms of censorship:

1. Right censoring refers to a situation in which the observed survival time of a patient exceeds the duration of the study or when the patient becomes unavailable for further follow-up during the course of the study. In this particular instance, it is established that the patient's survival was observed until a specific juncture, yet precise details regarding the timing of the event (in this case, the patient's demise) are unavailable.

2. Random censoring, also known as non-informative censoring, is based on the assumption that the individuals who are censored are selected randomly from the group at risk. This implies that the occurrence of censoring is independent of the event of interest.

The multiple myeloma dataset exhibits the presence of both types of censoring. For example, certain individuals within the patient cohort remain alive until the conclusion of the study, thereby constituting instances of right-censored observations. Additionally, there may be cases where patients have become unavailable for further monitoring, indicating occurrences of random censoring.

3.4.1 Survival Function

The 'Survival Function', abbreviated as 'S(t)', is the primary feature of survival analysis data. The likelihood that a subject (patient) survives from the time of origin (zero) to a given future time (t) is given by the survival function. In our context, it refers to the likelihood that a patient with multiple myeloma would live past a certain point after diagnosis. The following definition of the survival function (S(t)):

$$[St] = Pr(T > t)$$

where (T) is the survival time and (t) is the particular time under consideration. A declining step function is provided by (S(t)) that begins at 1 (signifying that all patients are alive at the beginning) and ends at 0 (stating that all patients have either experienced the event or have been censored). Understanding the distribution of survival times and the danger of the event over time requires an understanding of the survival function.

Survival analysis enables precise time-to-event data analysis and interpretation, particularly when working with censored observations. It offers a reliable way to calculate the odds of surviving, comprehend the variables affecting survival, and contrast the survival rates of various populations or treatment regimens. We can better understand the survival patterns and influencing factors for patients with multiple myeloma by using survival analysis techniques, such as the Kaplan-Meier estimate and Cox Proportional Hazards Model, with our multiple myeloma dataset.

3.4.2 Estimated via Kaplan-Meier

The survival function can be estimated from time-to-event data using the widely used non-parametric statistic known as the Kaplan-Meier (KM) estimate, also referred to as the product limit estimate. 'Non-parametric' in the context of survival analysis refers to the fact that the Kaplan-Meier estimate makes no assumptions regarding the statistical distribution of the survival times. Instead, it builds an empirical approximation of the survival function using the observed data.

The Kaplan-Meier technique calculates the odds of survival at each specific moment in time when an event takes place. The overall survival probability at each time point is then calculated by multiplying these probabilities together. This procedure yields a stepwise function that lowers at the time of each event and holds steady in between occurrences. The Kaplan-Meier survival curve has'steps' that represent events (or fatalities in our case), and 'plateaus' between steps that reflect times when there are no events.

The Kaplan-Meier method's estimated survival function, S(t), is given by:

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where di is the number of events at time ti,ni is the number of people at risk (who have not yet experienced the event or have been censored) at time ti, and ti are the individual separate event times.

The Kaplan-Meier estimate offers a visual summary of the patient survival statistics in the context of the multiple myeloma dataset. Since we begin by assuming that every patient is still alive, the survival probability is 1. The survival probability drops with time when events (deaths) take place, creating steps on the Kaplan-Meier curve. This gradual decline in the likelihood of survival indicates the events' overall effect.

3.4.3 Interpreting the Kaplan-Meier Survival Curve

An effective technique for visualising patient survival over time is the Kaplan-Meier survival curve. The survival probability is plotted on the y-axis against time (counting backwards from the diagnosis) on the x-axis.

The Kaplan-Meier curve begins at a survival probability of 1 in the context of the multiple myeloma dataset, signifying that all patients are alive at the beginning of the observation period. As time goes on, the curve starts to incline, signifying the occurrence of the important event—in this case, death. One or more events (deaths) correspond to each drop in the curve, and the size of the drop is proportional to the number of occurrences in relation to the number of people who are still at risk for the event at that time.

When analysing the Kaplan-Meier survival curve, it is important to keep in mind the following:

1. Rate of Events: Information about the rate at which events happen can be gleaned from the Kaplan-Meier curve's slope. A steeper drop in the curve indicates a higher occurrence rate, which denotes a poorer chance of surviving within that specific time period. A flatter area of the curve, on the other hand, denotes a time period with fewer incidents and a better likelihood of survival.

2. Survival Probability: The survival function's value at any given time point on the curve indicates how likely it is that a patient will live past that point. A patient has a 50% chance of living past 12 months after diagnosis, for instance, if the curve dives to a survival probability of 0.5 at that time.

3. Non-Increasing Function: Because the Kaplan-Meier curve is a non-increasing function, it only gets smaller or stays stable with time rather than growing. This trait represents the fact that the likelihood of survival does not rise throughout time.

4. Confidence Intervals: Around the survival curve, the Kaplan-Meier estimate frequently displays confidence intervals as shaded areas or delimited lines. These intervals offer a range of numbers that, with a certain degree of confidence, represent the genuine survival chance. More uncertainty in the survival estimate is indicated by a broader confidence range.

5. Comparing Groups: The Kaplan-Meier survival curve is particularly helpful for contrasting the chances of survival among various patient groups. We can visually compare the survival probabilities of different groups and determine whether some groups have better or worse survival outcomes by plotting individual Kaplan-Meier curves for each group (e.g., patients who participated in clinical trials versus those who did not, or male patients versus female patients).

Analysing the Kaplan-Meier survival curve requires comprehending the features and trends of the survival probabilities over time, the frequency of occurrences, and the comparison of survival probabilities between various groups. We can gain important insights into patient survival after diagnosis and how it may be influenced by numerous characteristics like age, gender, or participation in clinical trials by using this tool on our multiple myeloma dataset.

3.5 Cox Proportional Hazards Model

The Cox Proportional Hazards Model, also known as the Cox model, is a key contribution to the field of survival analysis. In contrast to conventional regression models that forecast a numerical or categorical result, the Cox model forecasts the hazard, which represents the immediate risk of encountering the event of interest, at any specific time point. The objective of survival analysis in the context of multiple myeloma is to ascertain the expected duration of survival following diagnosis, as well as to identify the elements that may exert an influence on this timeframe. The

Cox Proportional Hazards Model is a statistical technique that can provide valuable insights to academics and physicians in addressing these inquiries.

Given a set of covariates x for an individual, the hazard function $h(t|x)$ at time t is modeled as:

$$h(t|x) = h_0(t) \cdot e^{\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p}$$

Where:

- $h(t|x)$ is the hazard function for an individual with covariates (x) at time (t).
- $h0(t)$ is the baseline hazard function, representing the hazard when all covariates are zero.

3.5.1 Hazard and Hazard Ratio

- **Hazard**: -Definition: The risk of experiencing an event at a certain moment (t) is referred to as the hazard, which is frequently denoted as (h(t)). The event in a survival study is frequently death, but it might also be any interesting occurrence, such a recurrence or the development of a disease.

- **Significance**: The risk provides insight into the factors that influence how an event occurs. A high risk of an occurrence occurring at a given moment is indicated by a high hazard at that time. It's important to remember that the hazard may alter over time as a result of shifting risk dynamics.

- **Hazard Ratio**: The hazard ratio (HR), which depicts the proportion of dangers in two groups, is a measurement of the effect size. The HR would quantify the relative risk of an event (like mortality) in treatment A compared to treatment B, for instance, if you had two treatments, A and B.An HR of 1 denotes that there is no difference in the dangers between the two groups. In the numerator group, an HR larger than 1 suggests a higher risk, whereas an HR less than 1 suggests a lesser risk.

3.5.2 Assumption of Proportional Hazards

1. **Definition:** The Cox Proportional Hazards model is supported by the Proportional Hazards Assumption. It is predicated on the idea that the danger functions for various groups change proportionally over time. This indicates that even if the absolute risk (hazard) increases, the relative risk (their hazard ratio) between two groups stays the same.

2. **Importance:** By guaranteeing that the influence of predictor factors (such as treatment or gender) on survival is constant throughout time, the assumption makes the model simpler. If the influence changed over time, it would not be able to sum up the effect of predictors with a single number (the hazard ratio).

3. **Testing the Assumption:** It is critical to determine whether the proportional hazards assumption is valid in real-world applications. If not, the Cox model's conclusions could

be deceptive. To evaluate this premise, many statistical tests and graphical techniques (such as visualising Schoenfeld residuals) might be used.

4. **Consequences:** It is possible that a variable's impact changes over time if the proportional hazards assumption is broken for that variable. To comprehend the shifting effects in such circumstances, researchers may need to stratify their study or use more complicated models.

Chapter 4

Results and Discussion

4.1 Key Determinants of Survival Rates

4.1.1 The Impact of Age on Survival

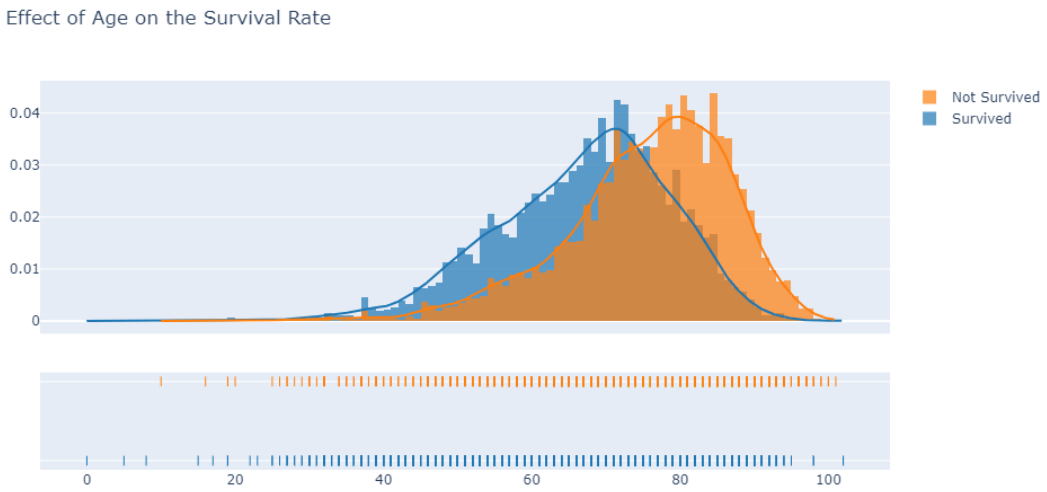

Figure 4.1: Effect of age on survival

The impact of age on survival reveals that individuals between the ages of 50 and 70 exhibit a comparatively higher rate of survival. The probability of mortality due to the condition is evident across various age groups, with the highest level of risk observed in individuals around their sixties. Nevertheless, it is worth noting that the likelihood of survival experiences a substantial decline after reaching the age of 80, indicating that advanced age constitutes a noteworthy risk element associated with diminished survival rates. The correlation between age and the probability of survival is illustrated in Figure 4.1.

4.1.2 Gender and Survival Rate

Approximately half of the overall population consists of males. Within this cohort of males, it was observed that around 30% managed to survive the ailment known as multiple myeloma, whereas the remaining 20% unfortunately did not survive. Conversely, the remaining portion of the population, constituting 50%, comprises individuals who identify as females. Within the cohort of female participants, approximately 35% exhibited survival throughout the duration of the study, while 15% experienced mortality as a result of the condition being investigated.

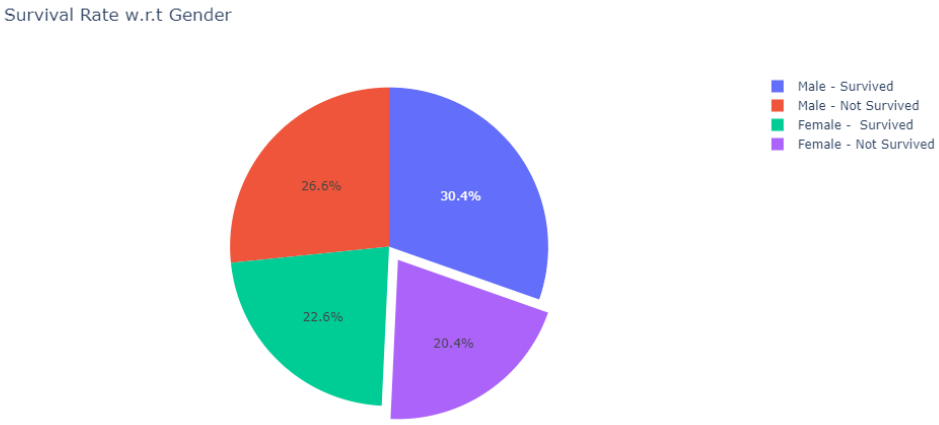

Figure 4.2: Survive rate w.r.t gender

The analysis of survival rates based on gender reveals a significant disparity between males and females. There is evidence suggesting that females exhibit a marginally elevated survival rate relative to males in the context of multiple myeloma. The gender disparity in survival outcomes is effectively depicted in figure 4.2

4.1.3 Clinical Trail Survival Rate

About 60% of the patients in the whole set of data, which is a big number, took part in a clinical trial. About 35% of these people with multiple myeloma made it through the disease, while the other 25% died from it. On the other hand, about 25% of the 40% of patients who did not take part in any clinical trials made it, while 15% did not.

In the end, 35% of the people who took part in clinical trials and had multiple myeloma survived, while 25% did not. Those who did not take part in clinical trials, on the other hand, had a 25% survival rate and a 15% death rate.

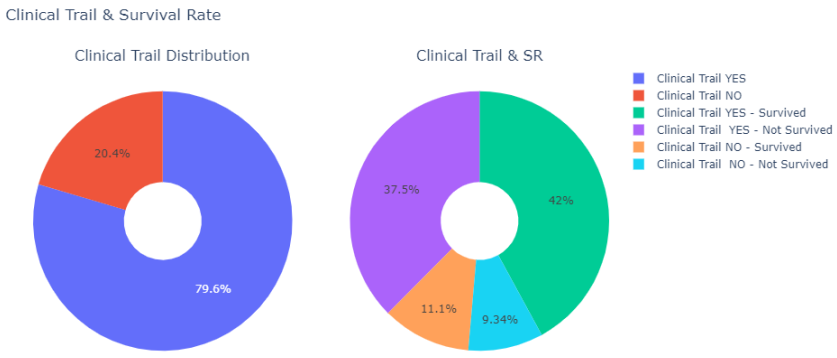

Figure 4.3: Clinical Trail Survival Rate

4.1.4 Comorbidity Score and its effects on the survival rate

The Comorbidity Score is an evaluative metric that serves as an indicator of the comprehensive well-being of an individual. The scoring system considers the presence of multiple comorbid conditions in patients, and this score has demonstrated significant predictive value for survival outcomes in individuals diagnosed with multiple myeloma. The data presented in the figure

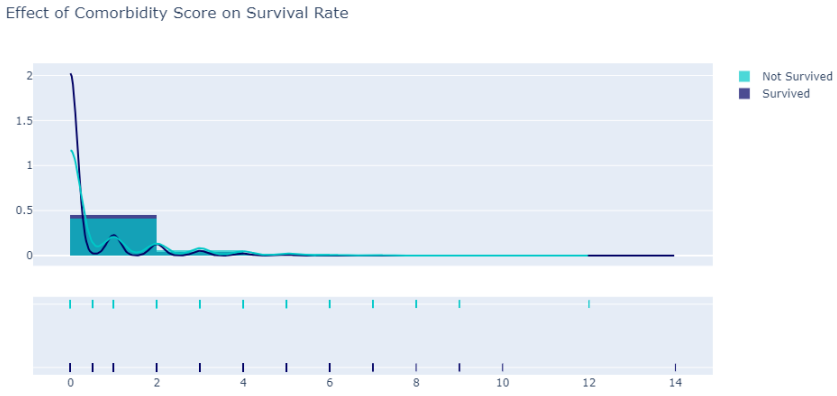

Figure 4.4: Effects on the survival rate

suggests a negative correlation between comorbidity score and survival rate, indicating that individuals with a higher comorbidity score experience a lower likelihood of survival. This finding aligns with the notion that individuals with comorbidities may face an elevated susceptibility to unfavourable consequences.

The distribution plot reveals that a significant proportion of the examined population exhibits a concentration of comorbidity scores within the lower to middle range. Nevertheless, it is worth noting that there exist individuals who exhibit elevated scores, suggesting the co-occurrence of

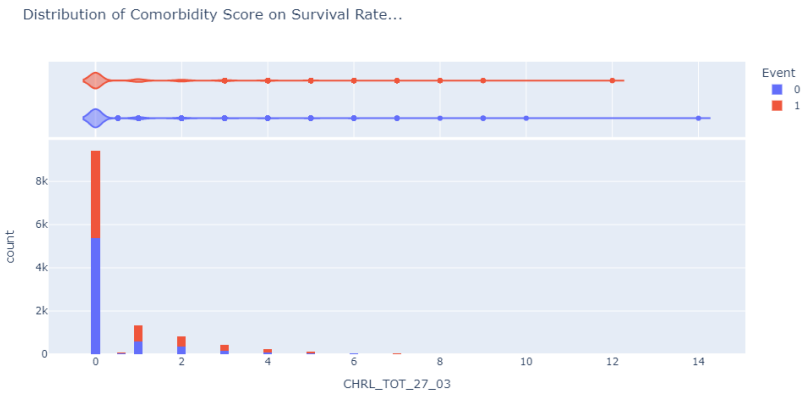

Figure 4.5: Distribution of Comorbidity Score

multiple comorbid conditions.

The predominant proportion of patients within this study cohort exhibits a comorbidity score that falls within the lower to middle spectrum. Nevertheless, there is a negative correlation between the comorbidity score and the survival rate, indicating that as the former increases, the latter tends to decrease. This observation implies that the presence of comorbidities may have an adverse effect on the prognosis of individuals diagnosed with multiple myeloma.

4.2 Survival Analysis by Different Groups

4.2.1 Survival Analysis by Age

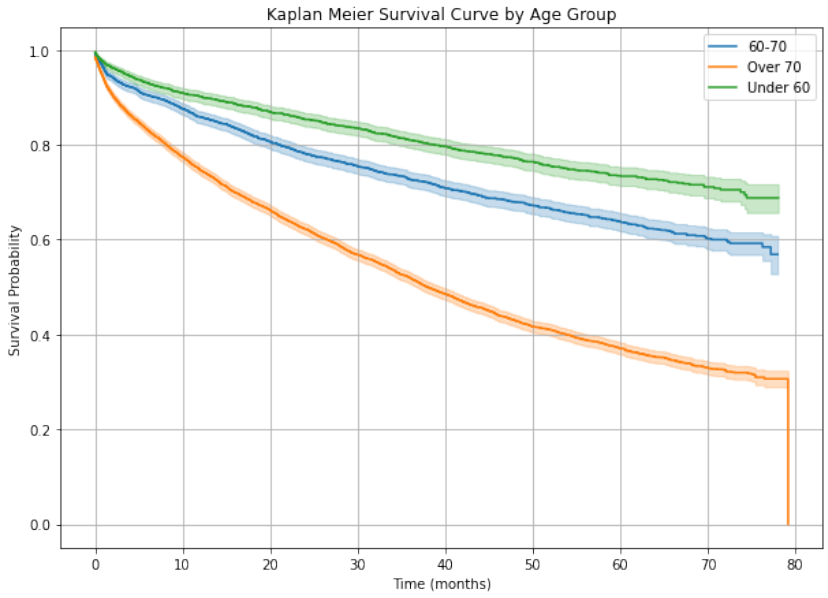

Figure 4.6: Kaplan Meier Survival Curve by Age Groups

In the present study on survival analysis, the patient cohort was stratified into three distinct age categories: individuals under the age of 60, those between the ages of 60 and 70, and individuals aged 70 and above. The objective of this classification was to analyse the potential influence of age at diagnosis on the duration of survival.

Survival curves using the Kaplan-Meier method were generated for each age group individually. The horizontal axis of the graph denotes the temporal progression (in months) subsequent to diagnosis, whereas the vertical axis signifies the estimated likelihood of survival at each corresponding time interval.

1. The survival curve for each age group commences with a survival probability of 1, signifying the presence of all patients at the inception of the study. As the passage of time unfolds, the likelihood of survival diminishes, indicating the manifestation of mortality within each respective age cohort.

2. Comparative Analysis of Age Groups: The utilisation of Kaplan-Meier survival curves enables us to visually assess and contrast survival probabilities among distinct age cohorts. The analysis of the plot reveals a notable decline in survival probability among individuals classified in the 'Over 70' group, in comparison to those in the 'Under 60' and '60-70' groups. These findings suggest that there is a correlation between older age (above 70 years) and a decreased likelihood of survival over time in patients diagnosed with multiple myeloma. Therefore, the age at which the disease is diagnosed can have a significant influence on the duration of survival in these patients.

3. Event Rate: The gradient of the survival curve for each age cohort offers valuable insights into the mortality rate. A more pronounced curve indicates an elevated mortality rate within the specified time period. An evident observation reveals a more pronounced decrease in survival rates within the 'Over 70' demographic, suggesting a heightened mortality rate among individuals in this age group.

   The utilisation of survival analysis based on age provides insights into the impact of patients' age at the time of diagnosis on their subsequent survival duration.

4.2.2 Survival Analysis by Comorbidity Score

In the survival analysis, patients were categorised according to their comorbidity scores, which are represented by the labels '0', '1', '2', '3', '4', and '5+'. These labels indicate the quantity of additional diseases (comorbidities) a patient has in addition to multiple myeloma. The Kaplan-Meier survival curves for each of these comorbidity categories have been plotted, allowing us to compare their survival outcomes.

1. Multiple distinct Kaplan-Meier survival curves, each representing a separate comorbidity group, are displayed on the graph. Let's attempt to interpret these analysis results.

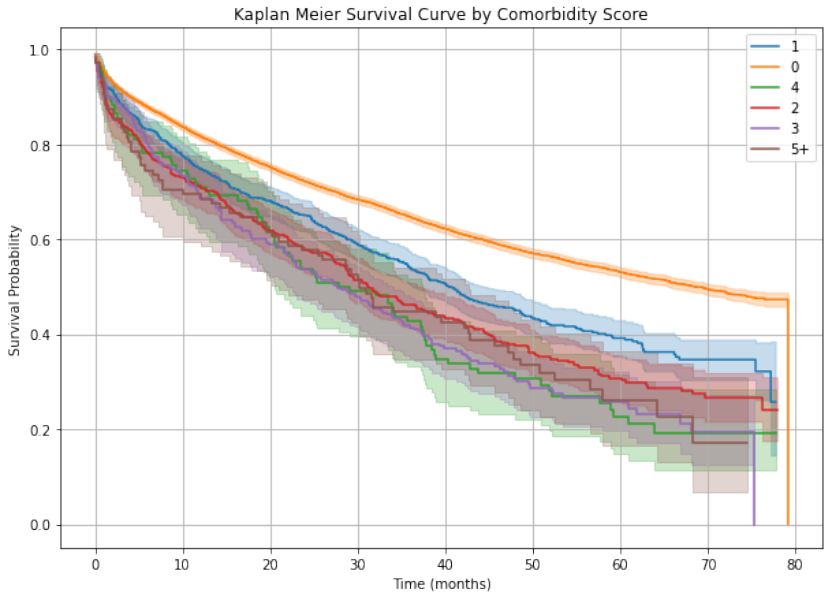

Figure 4.7: Kaplan Meier Survival Curve by Comorbidity Score

2. Survival Through Time: Each survival curve begins with a survival probability of 1, indicating that all comorbidity group patients are alive at the beginning of the study. As time passes, the probability of survival decreases due to the occurrence of fatalities within the group. The mortality rate at a particular time interval is greater the steeper the decline in the survival curve.

3. Comparative Analysis of Comorbidity Groups: The survival curves enable a visual comparison of survival probabilities across various comorbidity categories. Examining the survival curves at a specific time point, say 50 months, reveals that the survival probabilities for each comorbidity group vary. For instance, the survival probability of patients with a comorbidity score of '0' at 50 months appears to be around 0.75, indicating that a patient with no additional health conditions has a 75% chance of surviving beyond 50 months after diagnosis. The survival probability of patients with a comorbidity score of '5+' at the same time point is significantly lower, indicating a poorer prognosis for patients with multiple comorbid conditions.

4. The impact of comorbidities on survival: Generally speaking, we observe that the survival probability decreases as the comorbidity score rises. This suggests that having more comorbid conditions may have a negative impact on a patient's survival, potentially as a result of a cumulative effect on the patient's overall health and treatment responsiveness.

In the survival analysis by comorbidity score provides valuable insight into the impact of additional health conditions on the survival of multiple myeloma patients. Understanding this relationship can enlighten patient management and treatment strategies with the aim of enhancing patient survival.

4.2.3 Survival Analysis by BMI

In this Kaplan-Meier survival plot, patients are divided into four categories according to their Body Mass Index (BMI): ”Underweight,” ”Normal weight,” ”Overweight,” and ”Obese.” The graph illustrates how the survival probabilities of these groups fluctuate over time, allowing us to discern how a patient’s BMI at the beginning of the regimen may influence survival outcomes. Impact of BMI: The graph indicates that the BMI at the beginning of the treatment may affect

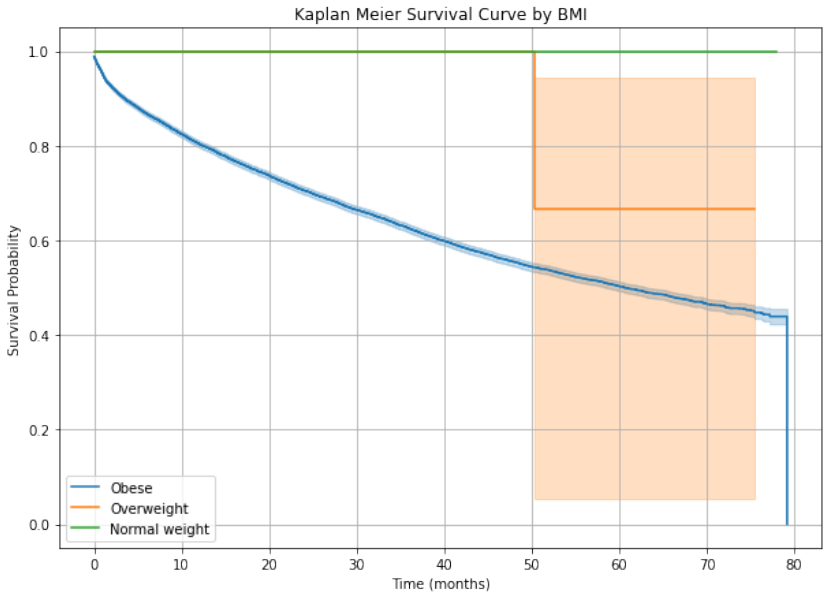

Figure 4.8: Kaplan Meier Survival curve by BMI

survival outcomes. Patients with ’Normal weight’ may have improved survival outcomes than those who are ’Underweight’, ’Overweight’, or ’Obese’. This could be due to a number of factors, including the patient’s overall health and physical fitness, the disease’s interaction with body obesity, and the patient’s resistance to treatment.

4.2.4 Survival Analysis by Time to First Treatment

The Kaplan-Meier survival analysis by Time to First Treatment plots survival probabilities for various groups of patients, categorised by the length of time between their diagnosis and the beginning of their first treatment.

The categories are ”1 month,” ”1-3 months,” ”3-6 months,” ”6-12 months,” and ”¿12 months.” This analysis enables us to examine how the length of time between the initiation of the first treatment and survival outcomes may vary.

Impact of Time to First Treatment: The graph indicates that the time between diagnosis and the initiation of the first treatment may have an effect on survival outcomes. Patients who begin treatment sooner after diagnosis, for instance, may have higher survival outcomes than those who begin treatment later. This could be due to a variety of factors, such as earlier intervention

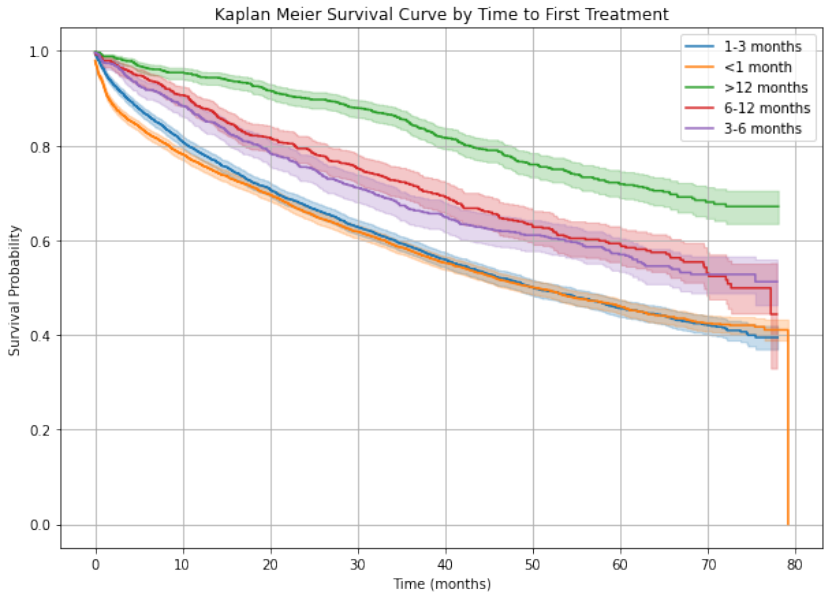

Figure 4.9: Kaplan Meier Survival curve by Time to First Treatment

slowing the progression of the disease or the patient's improved overall health at the onset of treatment.

4.2.5 Survival Analysis by Cancer Care Plan Intent

Curative, Non-curative, No active treatment, and Unknown are the categories represented by the Kaplan-Meier survival analysis by Cancer Care Plan Intent. This analysis enables us to evaluate the impact of the intended treatment strategy on survival outcomes and compare survival probabilities between these categories.

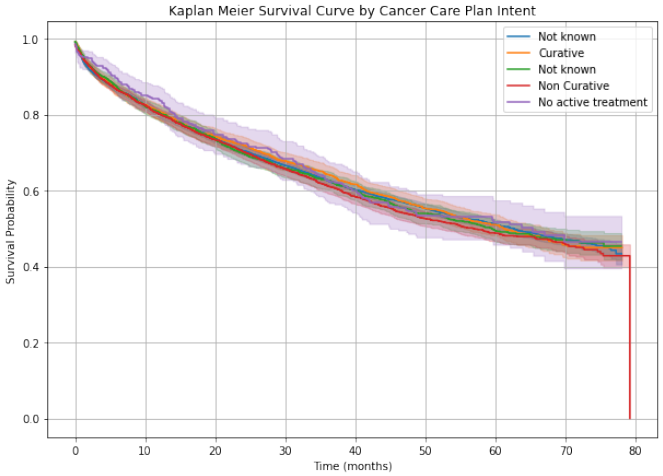

Figure 4.10: Kaplan Meier Survival curve by Cancer Care Plan Intent

At any given time point, it is possible to compare the survival probabilities of various cat-

egories. Upon examining the 50-month survival curves, we observe that survival probabilities vary between groups. Let's presume that the probability of survival at 50 months for patients with a curative intent is approximately 0.80. This suggests that a patient with a curative care plan has an 80 percent probability of surviving beyond 50 months after diagnosis. Depending on the intended treatment strategy, the survival probabilities of patients with various care plan intentions at the same time point may differ, indicating varying survival outcomes.

Effect of Care Plan Objectives: The overall conclusion drawn from the plot is that the intended treatment strategy may affect survival outcomes. Patients with a curative intent, for instance, may have improved survival outcomes than those with a non-curative intent or no active treatment. This could be due to a number of factors, including the efficacy of the treatments, the progression of the disease, or the patient's overall health and resilience.

4.2.6 Survival Analysis by Gender

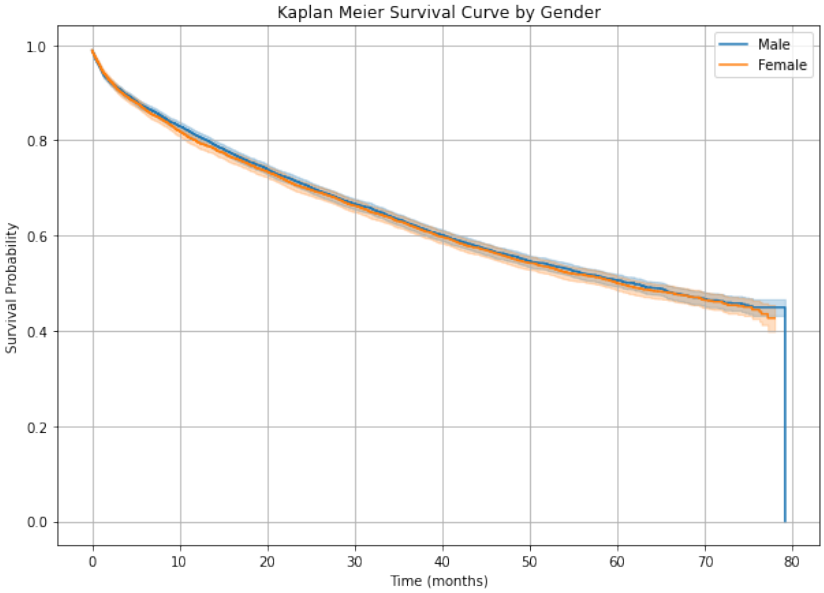

Figure 4.11: Kaplan Meier Survival curve by Gender

The Kaplan-Meier survival curves for male and female patients with multiple myeloma in the multiple myeloma dataset indicate a disparity in gender-based survival outcomes. This could suggest that gender is a significant factor influencing survival in multiple myeloma, with female patients exhibiting marginally higher survival probabilities than male patients.

4.3 Cox Proportional Hazards Model

4.3.1 Cox PH Fitter

The researchers utilised the Cox Proportional Hazards regression model to examine the impact of several predictors on the survival duration of patients in the multiple myeloma dataset.

```
CoxFitter.fit(df, duration_col='SurvivalTime',
              event_col='Event',
              show_progress = True)

Iteration 1: norm_delta = 0.56372, step_size = 0.9500, log_lik = -53016.59028, newton_decrement = 972.12055, seconds_since_star
t = 1.3
Iteration 2: norm_delta = 0.13071, step_size = 0.9500, log_lik = -52028.20972, newton_decrement = 46.76984, seconds_since_start
= 1.6
Iteration 3: norm_delta = 0.01352, step_size = 0.9500, log_lik = -51980.15904, newton_decrement = 0.54584, seconds_since_start
= 1.9
Iteration 4: norm_delta = 0.00028, step_size = 1.0000, log_lik = -51979.60878, newton_decrement = 0.00015, seconds_since_start
= 2.3
Iteration 5: norm_delta = 0.00000, step_size = 1.0000, log_lik = -51979.60863, newton_decrement = 0.00000, seconds_since_start
= 2.7
Convergence success after 5 iterations.

<lifelines.CoxPHFitter: fitted with 12492 total observations, 6622 right-censored observations>
```

Figure 4.12: Summary of Coxfitter

1. **Dimensions of the Dataset**: - The model underwent training using a dataset including 12,492 observations, each representing a distinct patient. The study utilised data from a total of 12,492 patients in order to examine the potential influence of various characteristics on their respective survival durations.

2. **Right-Censored Data**: - Among the total sample of 12,492 patients, a total of 6,622 patients were classified as right-censored. Right-censoring is a phenomenon that arises when the occurrence of the event of interest, such as death, is not seen for a specific patient during the designated research duration. There are several potential reasons for the absence of an event in the data. One possibility is that the event has not yet taken place, as the observation period may not have elapsed. Another possibility is that the patient in question withdrew from the study, so precluding the occurrence of the event. Lastly, it is also plausible that the study itself concluded before the event could be observed for the patient.

3. **Convergence of the Model**: - The optimisation process of the model achieved convergence after a total of five iterations. The occurrence of convergence is regarded as a positive indication, since it signifies that the model has successfully identified the optimal parameters that best align with the provided dataset. This suggests that the Cox model effectively captured the associations between the predictors and the survival periods in the iterations.

4.3.2 Goodness of Fit

The coefficients are crucial to comprehending patient survival when using the Cox Proportional Hazards model to examine Multiple Myeloma. These coefficients shed light on the log hazard ratios connected to numerous factors, like age, illness stage, or medical interventions. A positive

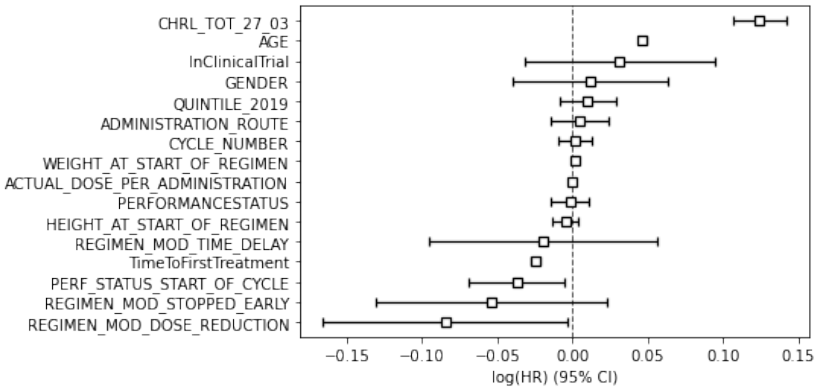

Figure 4.13: Visual representation of the log hazard ratios

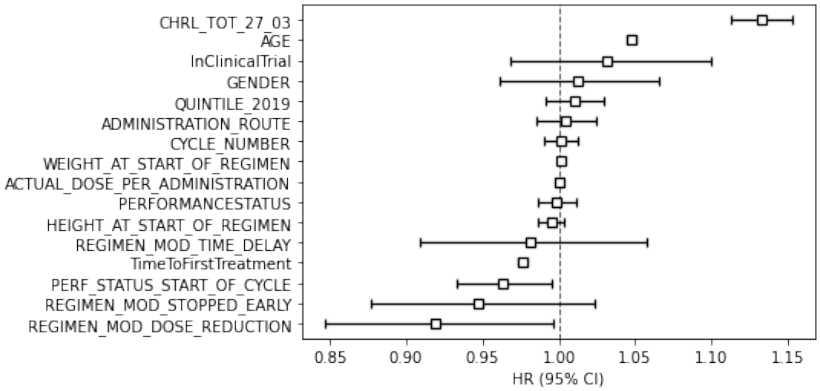

Figure 4.14: Visual representation of the hazard ratios

coefficient for a predictor, such as age, indicates that the risk or hazard of an event (such as the development of a disease or mortality) increases with each unit rise in age. A negative coefficient, on the other hand, denotes a protective effect or a reduced risk.

In reality, a positive coefficient for a particular treatment regimen may show that it is linked to a higher risk of illness progression than alternative treatments. Such findings can affect clinical trial design, additional research, and treatment decisions, making them essential for both physicians and researchers.

4.3.3 Statistical Significance and Its Importance

It's critical to identify which factors are statistically significant for the routine data on Multiple Myeloma patients. This knowledge can be obtained from the confidence intervals corresponding to each coefficient. A predictor may be statistically significant in changing the hazard at a 5% significance level if its confidence interval does not include zero. For instance, a medicine or treatment plan may be a great candidate for additional research or possibly a clinical trial if the confidence interval does not span zero.

Regular Multiple Myeloma data has great promise, both in terms of volume and the depth of insights it can provide when combined with reliable statistical techniques like the Cox model. The medical community may make educated judgments, customize therapies, and ultimately strive for better patient outcomes by identifying the variables that have a major impact on patient survival and disease progression. Such models and the insights obtained from them constitute the cornerstones of evidence-based medicine in the changing environment of cancer care, where individualized treatments are becoming the standard.

The Cox Proportional Hazards model provides crucial insights into the important indicators of patient survival in the context of examining the possibility of routinely gathered Multiple Myeloma cancer data. The risk associated with the disease's course and eventual results is largely determined by the predictors with 95% confidence intervals that do not include zero. The model has yielded the following conclusions:

1. Age as a Key Predictor

   - Confidence Interval: (0.044, 0.049)

   - Age stands out as a significant factor affecting patient outcomes, according to this insight. The risk of negative outcomes, such as mortality from multiple myeloma, increases as one gets older. This is consistent with the general concept of oncology that age frequently corresponds with the severity and prognosis of the disease.

2. Clinical Measurement (CHRL TOT 27 03)

   - Confidence Interval: 0.107 to 0.142

   - The variable emerges as a significant predictor, maybe corresponding to a particular clinical score or measurement. A higher score indicates a higher risk, suggesting the need

to monitor and maybe control this factor in normal care.

3. Performance Status at the Start

   - Confidence Interval (0.069, 0.005)

   - The risk is inversely influenced by the performance status, which may be a gauge of a patient's general wellbeing and capacity at the start of treatment. An improved performance status shows a strengthened resistance to the disease's negative effects, highlighting the significance of total patient health.

4. Treatment Modulation(REGIMEN MOD DOSE REDUCTION)

   - Confidence Interval: (-0.166, -0.004)

   - Modulating the therapy dose, particularly by reducing it, has a good effect on survival. This may highlight the value of individualized treatment strategies that take the patient's health profile and stage of the disease into account.

5. Promptness in Treatment Initiation(TimeToFirstTreatment)

-Confidence Interval: (-0.028, -0.021)

- The length of time between diagnosis and the start of the first treatment has a substantial impact on survival rates. A protracted period could be a sign of underlying problems with the healthcare system or the patient's health, requiring prompt actions.

These important predictors highlight the enormous potential of frequently gathered Multiple Myeloma data. Healthcare providers can customize treatments, prioritise interventions, and ultimately strive for improved patient outcomes by recognizing and utilizing these factors. The importance of such routine data in guiding evidence-based, individualized oncology care is reiterated by this study.

4.3.4 Assessing the Predictive Strength of Variables in Routine Multiple Myeloma Cancer Data

Understanding the statistical significance of numerous parameters and their predictive value in determining patient outcomes is crucial to our investigation of the often obtained Multiple Myeloma cancer data. In this aspect, the p-values are an accurate indicator. With a p-value of less than 0.0005, age stands out as a significant predictor, indicating a strong statistical link between growing older and a higher risk of dying from multiple myeloma. A clinical score might be represented by the variable "CHRL TOT 27 03," which has a p-value of less than 0.0005. This variable has a tremendous amount of predictive power. Statistics show that a rise in this score is linked to a higher probability of passing away. Another crucial finding is the importance of the amount of time that passed prior to the start of treatment. Statistics show that delaying treatment initiation is associated with a lower risk, as shown by the p-value of less than 0.0005. Contrarily, as shown by their high p-values, several variables, such as performance

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
GENDER	0.012	1.012	0.026	-0.040	0.064	0.961	1.066	0.000	0.455	0.649	0.623
AGE	0.046	1.048	0.001	0.044	0.049	1.045	1.050	0.000	36.890	<0.0005	987.198
QUINTILE_2019	0.010	1.010	0.010	-0.009	0.029	0.991	1.029	0.000	1.061	0.289	1.793
PERFORMANCESTATUS	-0.002	0.998	0.006	-0.014	0.011	0.986	1.011	0.000	-0.256	0.798	0.325
CHRL_TOT_27_03	0.124	1.132	0.009	0.107	0.142	1.113	1.153	0.000	13.803	<0.0005	141.553
HEIGHT_AT_START_OF_REGIMEN	-0.005	0.995	0.004	-0.013	0.004	0.987	1.004	0.000	-1.083	0.279	1.842
WEIGHT_AT_START_OF_REGIMEN	0.001	1.001	0.001	0.000	0.003	1.000	1.003	0.000	2.185	0.029	5.114
CYCLE_NUMBER	0.002	1.002	0.006	-0.009	0.013	0.991	1.013	0.000	0.287	0.774	0.370
PERF_STATUS_START_OF_CYCLE	-0.037	0.964	0.016	-0.069	-0.005	0.933	0.995	0.000	-2.272	0.023	5.438
REGIMEN_MOD_DOSE_REDUCTION	-0.085	0.919	0.041	-0.166	-0.004	0.847	0.996	0.000	-2.045	0.041	4.613
REGIMEN_MOD_TIME_DELAY	-0.019	0.981	0.039	-0.095	0.056	0.910	1.058	0.000	-0.501	0.616	0.698
REGIMEN_MOD_STOPPED_EARLY	-0.054	0.947	0.039	-0.131	0.023	0.877	1.023	0.000	-1.379	0.168	2.575
ACTUAL_DOSE_PER_ADMINISTRATION	-0.000	1.000	0.000	-0.000	0.000	1.000	1.000	0.000	-0.109	0.913	0.132
ADMINISTRATION_ROUTE	0.005	1.005	0.010	-0.015	0.024	0.985	1.024	0.000	0.473	0.636	0.653
InClinicalTrial	0.031	1.032	0.032	-0.032	0.095	0.969	1.099	0.000	0.971	0.332	1.592
TimeToFirstTreatment	-0.024	0.976	0.002	-0.028	-0.021	0.973	0.979	0.000	-14.833	<0.0005	162.932

Figure 4.15: Result of Cox Model

status, gender, cycle count, and the actual dose per administration, don't seem to hold the same significance in predicting the hazard. Given their substantial standard errors and expansive confidence intervals, these variables may not be particularly important in determining the death occurrences.

4.3.5 Harnessing the Power of Routine Data in Multiple Myeloma Patient Outcomes

Understanding the Hazard Ratio (HR) emerges as a crucial stage in the process of learning about the complexities of Multiple Myeloma via the lens of regularly collected data. The HR, which represents the risk shift brought on by a one-unit change in a particular factor, acts as a compass to lead us through the complex world of cancer prognosis.

We discover crucial insights by delving deeply into the data of multiple myeloma patients. Notably, despite gender showing a slight 1.2% difference in risk, this illustrates how even tiny factors can have an impact on results. Males have a slightly higher risk than females, but the difference isn't statistically significant, as shown by the 95% confidence interval that surrounds the data. Age, a common indicator of health problems, is still important in this situation. The mortality risk increases noticeably by 4.8% for each additional year a patient lives.

Another measurement, "CHRL TOT 27 03," which may be a clinical test, shows promise as a predictor. This score increased at the same time that risk increased by 13.2%, underscoring its clinical significance. Mortality has an inverse relationship with the patient's performance status at the start of the treatment. An improvement in this condition coincides with a 3.6% decrease in the probability of dying, highlighting the importance of holistic health.

A further benefit of treatment changes, particularly dose decreases, is the 8.1% decrease in associated hazards. This result highlights the therapeutic potential of customized treatment plans. The importance of fast treatment initiation serves as the study's clincher observation, underlining the necessity of timely medical action. The mortality risk is reduced by 2.4% for every month that treatment is delayed, highlighting the need of early diagnosis and intervention.

4.3.6 Deciphering Predictive Potency through Concordance in the Realm of Multiple Myeloma Data

The metric of Concordance, often known as the c-index, emerges as a brilliant thread displaying the model's capacity for prediction in the complex web of survival analysis. This metric, which is based on the basic tenet of ordering events, assesses the model's skill at accurately predicting the chronological order of events. Imagine a situation in the dataset where two different multiple myeloma patients, A and B, are being examined. The Concordance indicator measures the likelihood that our analytical model will correctly predict that patient A is at a higher risk than patient B if empirical data shows that patient A's event occurred before that of patient B.

The c-index provides an insightful indicator of model effectiveness by navigating the numerical spectrum between 0 and 1. A midpoint value of 0.5 suggests that a model's predictive power is equal to that of pure randomness, making it useless. In contrast, a pinnacle score of 1 represents the gold standard in survival models for flawless predicted discrimination.

The model's c-index, which delves deeply into the frequently gathered Multiple Myeloma data, reveals a value of 0.668. This numerical finding, which goes beyond mere chance, indicates that our model does a good job of differentiating across patients based on their event timelines. This c-index, which is 0.668, reflects the model's skill in determining the temporal order of events as well as in enhancing the ability of routine data to create reliable predicting frameworks for the outcomes of multiple myeloma patients.

4.3.7 Partial Akaike Information Criterion (AIC) in the Landscape of Multiple Myeloma Data

In the immense ocean of regularly gathered data on multiple myeloma, the Akaike Information Criterion (AIC) stands out as a light of direction when delving into the complex world of model selection. AIC, a renowned statistical measure, carefully evaluates how faithfully the model replicates the underlying data. While doing so, it constantly keeps an eye on the delicate balance between the model's inherent complexity and the amount of factors it can control. With careful calibration, the twin goals of achieving data representation and avoiding overfitting are both achieved. In summary, those wielding lower AIC values frequently gain the throne of greater data fit when faced with a multitude of models competing for dominion on a dataset.

Our exploration of the multiple myeloma data, however, leads us to a more complex version of this metric called the "Partial Akaike Information Criterion (AIC)". This improved AIC incarnation incorporates partial likelihoods and is tailored specifically for the Cox Proportional Hazards models, providing a more contextualized viewpoint. We are left with a Partial AIC score of 103991.217 for the dataset at hand. This numerical evidence acts as both a compass and a snapshot of the model's suitability. If our analytical odyssey were to explore different model terrains or nuanced specifications, their respective Partial AICs would prove to be invaluable allies in helping us identify the model that successfully combines parsimony with accuracy and accurately captures the potential and nuances of the routine multiple myeloma data.

4.3.8 Understanding the Partial Effects of Covariate Roles in Multiple Myeloma Survival Stories

Entering the huge world of survival analysis, especially when dealing with ordinary multiple myeloma cancer data, frequently necessitates a detailed investigation of the partial impacts of the covariates. We explore the complex landscapes of survival curves in order to understand how the survival trajectories are affected when one covariate is changed while the others remain constant. This introspective investigation opens doors to see the delicate dance between certain variables and survival outcomes, illuminating their connections in subtle ways.

Age, ComorbidityGroup, and TimeToFirstTreatment have taken center stage among the countless covariates that make up our dataset, all of which have p-values ¡ 0.0005. Their statistical significance suggests a strong connection to the tragic occurrence of death, highlighting their key role in the survival story. Additionally, although its impact is more minor, the performance status at the start of the treatment cannot be disregarded. It also joins the list of significant variables with a p-value of 0.023, underscoring the deep relationships it has with survival results.

Partial Effects of Varied Time To first treatment Groups

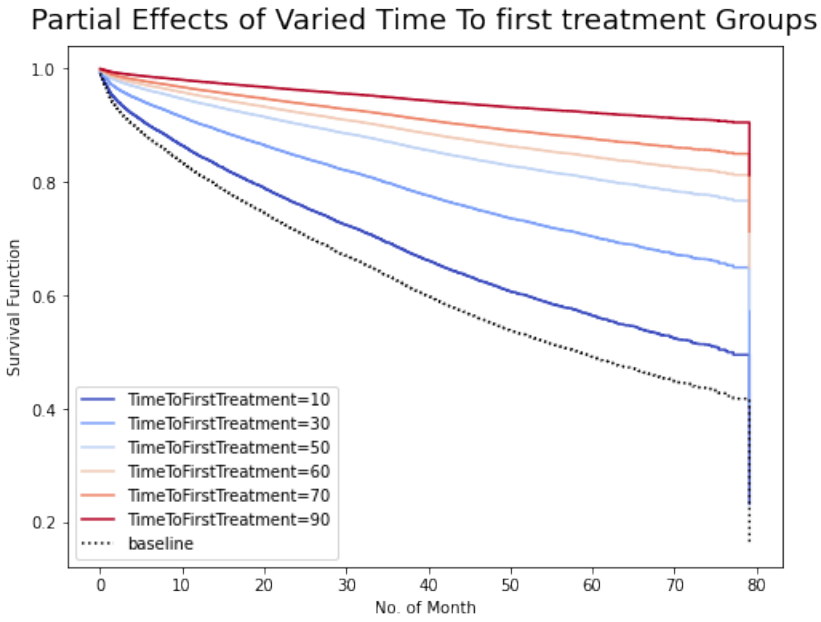

Figure 4.16: Partial Effects of Varied Time To first treatment

The graph clearly illustrates that patients who initiate their treatment within 10 months of diagnosis exhibit a greater likelihood of survival over time in comparison to individuals who postpone their treatment. The survival outcome remains the most favourable over the course of 10 months, as evidenced by the highest point on the curve along the time axis.

The detrimental impact of delay: With an increase in the time to initial treatment from 30

to 90 months, there is a corresponding decrease in the probability of survival. The downward trend of the curves becomes more pronounced as the delay in treatment initiation increases. As an example, it is observed that the survival curve of patients who initiate treatment following a 90-month delay after diagnosis exhibits the lowest rates. This finding emphasises the adverse consequences associated with a protracted delay in treatment initiation.

Middle Ground: Patients who begin their treatment around 50-60 months after diagnosis have intermediate survival outcomes, better than those who wait for 70 months or more but worse than those who start within 30 months.

Partial Effects of Varied Age Groups

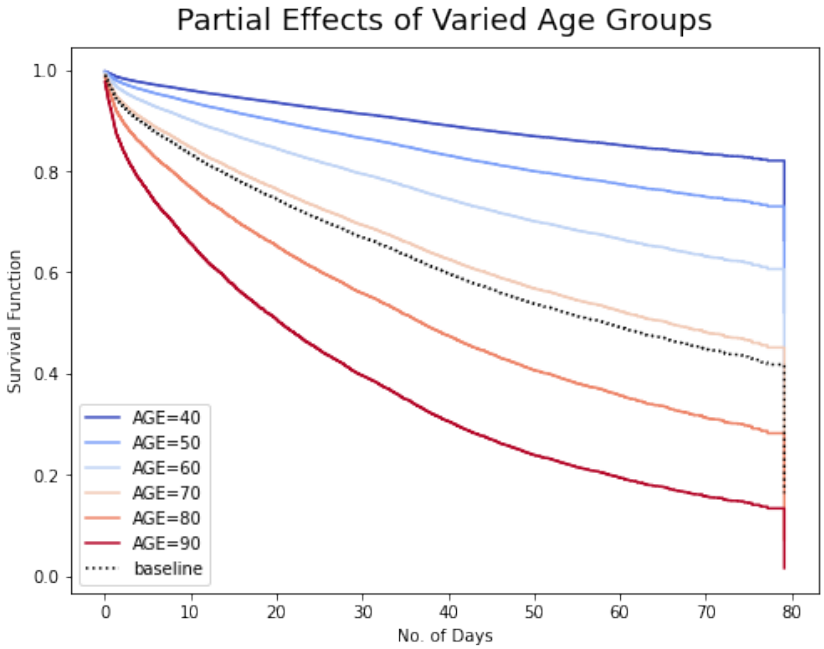

Figure 4.17: Partial Effects of Varied Age

The survival curves illustrate the probabilities of survival for patients in various age groups, spanning from 40 to 90 years old, as they advance through the duration of time (measured in months) following their diagnosis.

The survival curve exhibits the highest rates of survival consistently throughout the entire time axis for patients in the younger age group of 40 years old. This finding illustrates that individuals in the younger age bracket, particularly those approximately 40 years old, exhibit the greatest likelihood of survival in comparison to other cohorts. The resilience of their bodies and their potential for expedited recovery may be factors that contribute to this observed pattern.

The phenomenon of ageing is associated with a decrease in the likelihood of survival. The survival rates of patients aged 50, 60, and 70 exhibit a gradual decline, indicating a negative correlation between age and survival probability. The observed decrease can be attributed to

various factors, such as a potentially diminished immune response, the existence of additional comorbidities, and a generally diminished capacity to withstand intensive therapeutic interventions.

Partial Effects of Varied Comorbidity Group Values

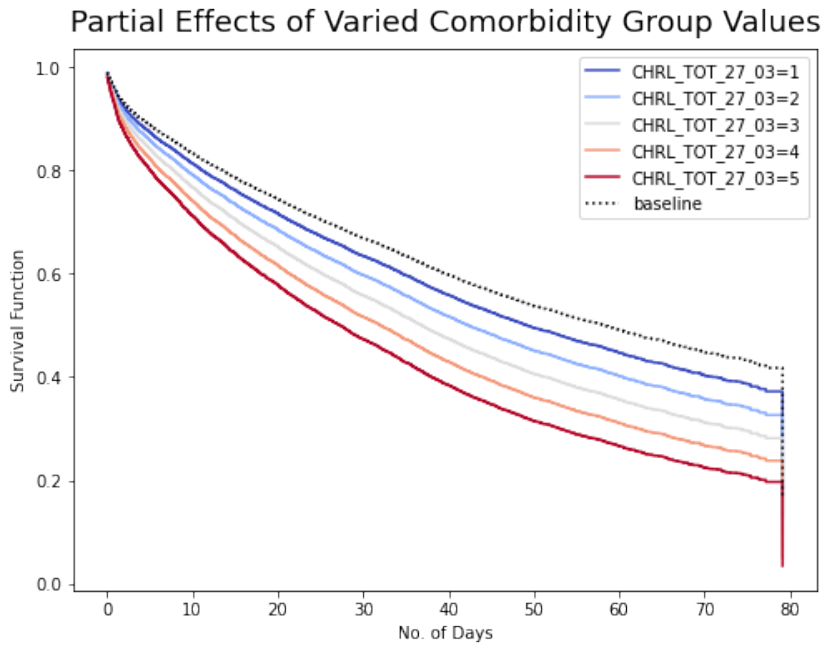

Figure 4.18: Partial Effects of Varied Comorbidity

The survival probability of elderly patients exhibits a significant decline, particularly among individuals aged 80 and 90 years. The survival curves of elderly patients exhibit the lowest values, indicating a heightened risk and diminished likelihood of survival.

The data presented in the plot clearly demonstrates a negative correlation between the variable 'CHRL TOT 27 03' and survival probability, indicating that higher values of 'CHRL TOT 27 03' are associated with lower chances of survival. The curve representing patients with a score of 1 exhibits the greatest elevation throughout the entire temporal axis, implying that patients with a lower score experience a more favourable likelihood of survival.

The probability of survival exhibits a consistent decrease as scores range from 1 to 5. The survival probability demonstrates a gradual decrease in patients with scores of 2, 3, and 4, as evidenced by the observed curves. This observation suggests that there is a progressive decrease in the likelihood of survival as the clinical measurement or score increases. The most significant decrease observed was for a score of 5. The survival probability of patients with a score of 5 exhibits the most substantial decline, indicating that this group is at the greatest risk compared to the other groups. The survival probability of this particular group is significantly lower in comparison to the other groups, particularly when compared to patients who have a score of 1.

Partial Effects of Varied PERF STATUS START OF CYCLE

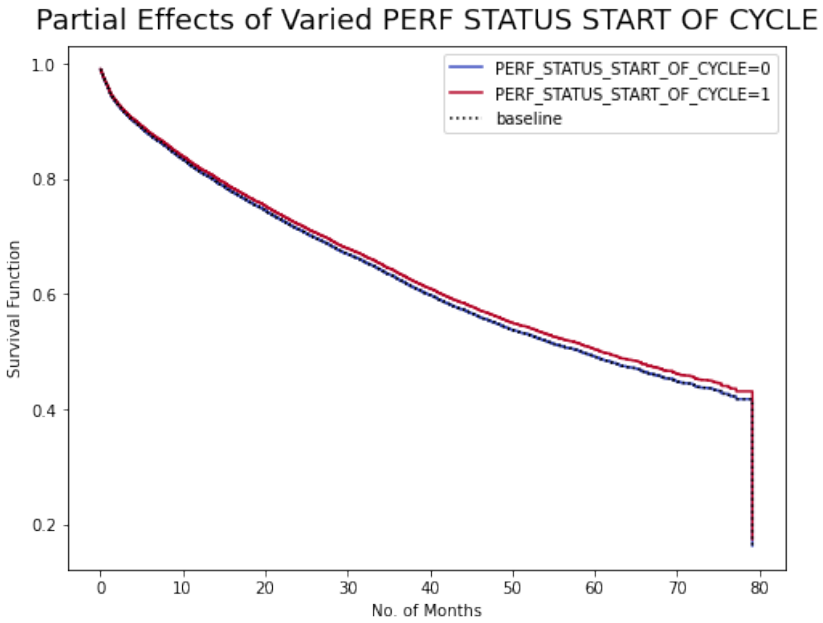

Figure 4.19: Partial Effects of Varied PERF STATUS START OF CYCLE

The plot clearly demonstrates that patients with a 'PERF STATUS START OF CYCLE' value of 0 exhibit a greater likelihood of survival compared to those with a value of 1. The curve representing the former consistently maintains a higher position than the curve representing the latter throughout the observed period.

Immediate Decline for State 1: The decline in the curve representing patients with a value of 1 is more prominent in the early stages, indicating a significantly elevated risk during the initial months. This suggests that individuals undergoing treatment may encounter immediate difficulties or complications following the initiation of their treatment regimen. Stability following Initial Decline: Although both groups exhibit a decrease in survival probabilities over time, the rate of decline reaches a stable state for both cohorts subsequent to the initial months. Nevertheless, there continues to be a disparity in survival probabilities between the two cohorts.

4.3.9 The Illuminating Power of the Log-Likelihood Ratio Test

The Log-Likelihood Ratio Test acts as a sentinel in the maze of statistical hypothesis testing, evaluating the compatibility of two different models based on their goodness of fit. Its main goal is to compare the effectiveness of our current model to a less sophisticated alternative, frequently a null model devoid of predictors. The result is a test statistic that fits a chi-square distribution. The statistic produced from our dataset, clocking in at 2073.963 over 16 degrees of freedom, emphasizes the current model's improved fit compared to its more straightforward version. This finding is amplified by the strong negative logarithmic p-value, which confirms

that the predictors we chose for the model are not just placeholders but have a significant impact on patients' survival trajectories when dealing with multiple myeloma. In essence, this test offers a strong validation of the importance our variables, which serve as beacons in the huge sea of ordinary multiple myeloma data, bring to the analytical table.

Chapter 5

Conclusion

5.1 From Raw Data to Revelations: A Myeloma Exploration

The exploration into the realm of multiple myeloma using routine data has been both enlightening and insightful. We began this journey with a primary objective: to untangle the myriad threads of data to derive meaningful patterns and relationships, with the ultimate aim of enhancing patient care and therapeutic outcomes.

- Untangling the Data Fabric: We were first given access to a wide tapestry of unstructured data. We processed and cleaned this data with great care, handling it like trained craftsmen to ensure its reliability and integrity. By laying the foundations for subsequent investigations, this initial stage made sure that the findings reached were supported by reliable and accurate data.
- Using the prepared data, we set out on a journey through the plethora of variables. Each factor, including age and clinical scores, served as a lighthouse, shedding light on particular aspects of multiple myeloma. We attempted to comprehend how each of these factors affected the course of the disease and the overall survival rates through statistical modeling and survival analysis.
- The discovery of the multifactorial nature of multiple myeloma was among the most profound ones made throughout this investigation. The complexity of the condition could not be adequately captured by a single variable. Age, gender, and comorbidity scores interacted in complex ways to provide a complex picture of the course of the disease and its results. In order to derive comprehensive insights, it was essential to recognize these interrelationships.
- The Power of Visual Representation: The Kaplan-Meier survival curves and other graphic representations were crucial to our trip. These visual aids not only made it easier to interpret complicated statistical results, but they also highlighted minute details and trends that could have gone unnoticed otherwise.

- The Development of Insights: Initial theories were tested, improved upon, or confirmed as we dug deeper into the data. What surfaced was a more complex, empirically based view of multiple myeloma. Our ideas progressed, reflecting the depth and breadth of our investigation, from comprehending the significance of early treatment commencement to identifying the small gender inequalities in survival rates.

5.2 Key Takeaways

5.2.1 Age as a Silent Protagonist

The research revealed that age is the single biggest factor affecting the likelihood of survival. The drop in survival rates for people above the age of 70 was particularly obvious. This isn't only a sign of aging; it could also point to age-related vulnerabilities like diminished physiological resilience, slowed recovery times, and greater susceptibility to adverse effects of medications. Because of this finding, treatment plans for older patients must be age-adapted and may even incorporate geriatric examinations.

5.2.2 The Gendered Dance of Survival

Our research revealed gender-based variations in survival outcomes that are small but noteworthy. On average, females had marginally higher survival rates. This may not only be a statistical outlier; it may also have biological, hormonal, or even treatment-related explanations. These findings call for more investigation into treatment approaches that are specific to gender and even the creation of remedies that are gender-specific.

5.2.3 Clinical Trials as Beacons

The data highlighted the significance and prospective benefits of clinical trial participation. Patients participating in clinical trials typically have access to the most recent therapeutic innovations, stringent monitoring, and a structured care protocol. These factors may have contributed to the increased survival rates of this cohort. This highlights the need for greater clinical trial advocacy, ensuring that patients are aware of and have access to these opportunities.

5.2.4 The Weight of Comorbidities

The data highlighted the significance and prospective benefits of clinical trial participation. Patients participating in clinical trials typically have access to the most recent therapeutic innovations, stringent monitoring, and a structured care protocol. These factors may have contributed to the increased survival rates of this cohort. This highlights the need for greater clinical trial advocacy, ensuring that patients are aware of and have access to these opportunities.

5.2.5 Timeliness in Treatment

The data confirmed emphatically the importance of early interventions. The sooner treatment is initiated following a cancer diagnosis, the better the prognosis for survival. This may be due to the fact that early-stage disease is more responsive to treatment, there are fewer complications, and the patient is in better health to begin with. The findings highlight the significance of early detection campaigns and streamlined care pathways to facilitate prompt interventions.

5.3 Implications for Clinical Practice

Analysis of routine Multiple Myeloma (MM) cancer data provides not only academic enrichment, but also insights that can directly influence clinical practices. This study's findings have the potential to alter patient care, approaches to therapy, and even policy decisions regarding multiple myeloma (MM).

- **Early Intervention Is Crucial:** Our analysis highlighted the significance of initiating treatment promptly. The survival rates of patients who began treatment promptly after diagnosis were superior to those who delayed treatment. Therefore, clinicians should prioritize early interventions for MM patients in order to improve their prognosis.

- **Comorbidity Management:** The research demonstrated a direct correlation between the comorbidity score and survival outcomes. Survival rates decreased as the score (indicating the number of co-existing conditions) increased. This insight is crucial for clinicians because it highlights the need for a holistic treatment approach in which not only MM but all of a patient's health conditions are actively managed and treated.

- **Age and Survival:** Age emerged as a significant factor influencing survival rates, particularly for patients older than 70. This finding can guide geriatric care for patients with multiple myeloma, leading to individualized treatment plans that take into account the specific challenges and requirements of older patients.

- **Gender Differences :** The disparity in gender-based survival suggests that biological differences between males and females may influence disease progression or treatment efficacy. Clinicians should be aware of these differences and may want to consider gender-specific monitoring or treatment protocols.

- **The Power of Clinical Trials:** Patients who participated in clinical trials had a distinct prognosis for survival than those who did not. This may indicate the potential benefits of innovative treatments or the need for more stringent surveillance in clinical trial settings. Clinicians should consider referring eligible patients to clinical trials, as it may provide them with more effective treatment options.

- **BMI's Role in Treatment:** The impact of BMI on survival rates suggests that a patient's physical health, nutrition, and endurance have an effect on their prognosis. It is possible that nutritional counseling, fitness programs, and other supportive therapies will become standard components of care for MM patients.

- **Tailoring Treatment Plans:** The data revealed that patients with a curative care plan intent had higher survival rates than those with a non-curative care plan intent or no active treatment. This emphasizes the significance of aggressive and targeted treatment approaches for enhancing patient outcomes.

The study has numerous consequences. Through the analysis of patterns in routine MM data, various areas in clinical settings that necessitate attention have been identified. By incorporating these discoveries into the realm of clinical practice, healthcare practitioners have the ability to increase treatment approaches, enhance patient guidance, and potentially augment survival rates for those diagnosed with MM. The integration of data-driven insights and clinical experience presents a potential for enhanced patient care in the field of multiple myeloma.

5.4 Future Directions

The study of routine cancer data, especially in the context of multiple myeloma, has yielded a number of findings and insights. While the current study has advanced our understanding of the potential of such data, the ever-evolving fields of healthcare, technology, and data analytics present numerous opportunities for future research and development. Listed below are some suggested future directions:

- **Integration with Other Data Sources**: - Future research could investigate incorporating routine cancer data with other data sources, such as genomics databases, electronic health records, and patient-reported outcomes. This would enable for a more robust and extensive data landscape, which could reveal deeper insights.

- **Advanced Analytical Methods**: - With the advent of machine learning and artificial intelligence, it is possible to employ more complex models and strategies. Deep learning, for example, may uncover non-linear patterns or interactions in the data that conventional methods may overlook.

- **Real-time Predictive Analysis:** - Based on predictive models, it is possible to develop systems that can analyze data in real-time and provide healthcare providers with immediate insights. This may be crucial for making prompt clinical decisions.

- **Customized Treatment Plans:** - Multiple myeloma patients have the potential to receive treatment plans tailored to their unique genetic makeup, health history, and other factors as we collect more data and improve our analytical methodologies.

- **Data Quality and Standardization:**- Future efforts could concentrate on improving the quality of data acquisition and promoting standardization across various datasets. This would ensure that data is consistent, trustworthy, and can be incorporated seamlessly across platforms.
- **Ethical and Privacy Issues:** - As personalized medicine and data analytics become increasingly prevalent, patient privacy and data security will become a greater concern. Future research could investigate methods for protecting data privacy while gaining insights from the data.
- **Collaborative Research:**- Multiple myeloma, like numerous other diseases, is a worldwide concern. There is potential for international collaborative research that combines data from various regions and populations in order to gain a deeper understanding of the disease.
- **Patient Engagement:** - Future research may also investigate methods to actively engage patients in their care. This may involve the development of tools or platforms that enable patients to access and even contribute to their data, fostering a more collaborative approach to healthcare.

While the present study has paved the way for understanding the potential of routine cancer data in the context of multiple myeloma, there is still much to discover. These future directions not only promise improved patient care, but also advancements in the field of oncological research and treatment as a whole.

5.5 Limitations

Despite being thorough, the study on the potential of routine Multiple Myeloma cancer data has a number of drawbacks. The synthetic Simulacrum dataset's temporal limits limit the scope to developments up to a specific time, and it might not capture all the subtleties of real patient data. Although the Cox Proportional Hazards Model is a robust methodology, it does include some specific assumptions that, if broken, can affect the outcomes. The analysis may have missed other important external influences because it concentrated on particular variables. The conclusions, which were drawn from a particular population dataset, might not be applicable to everyone. The huge dataset could still contain errors or missing data despite thorough data cleansing. The extent of insights was further constrained by ethical issues surrounding data modeled after actual data, the probabilistic nature of prediction models, technological limitations of the tools used, and the basic data science approach without further clinical integration.

Bibliography

David, A.R., & Zimmerman, M.R. (2010). Cancer: an old disease, a new disease or something in between? *Nature Reviews Cancer* **10**(10), 728–733.

Jensen, O.M., Parkin, D.M., MacLennan, R., Muir, C.S., & Skeet, R.G. (1991). Cancer Registration: Principles and Methods. International Agency for Research on Cancer.

Prokosch, H.U., & Ganslandt, T. (2009). Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods of Information in Medicine* **48**(1), 38–44.

Collins, F.S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine* **372**(9), 793–795.

Coveney, P.V., Dougherty, E.R., & Highfield, R.R. (2016). Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2080), 20160153.

Noone, A.-M., Cronin, K.A., Altekruse, S.F., et al. (2017). Cancer Incidence and Survival Trends by Subtype Using Data from the Surveillance Epidemiology and End Results Program, 1992-2013. *Cancer Epidemiology and Prevention Biomarkers* **26**(4), 632–641.

Smith, B.D., Bellon, J.R., Blitzblau, R., et al. (2018). Radiation therapy for the whole breast: Executive summary of an American Society for Radiation Oncology (ASTRO) evidence-based guideline. *Practical Radiation Oncology* **8**(3), 145-152.

Keegan, T.H.M., DeRouen, M.C., Press, D.J., Kurian, A.W., & Clarke, C.A. (2012). Occurrence of breast cancer subtypes in adolescent and young adult women. *Breast Cancer Research* **14**(2), R55.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., & Fotiadis, D.I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* **13**, 8-17.

Clark, T.G., Bradburn, M.J., Love, S.B., & Altman, D.G. (2003). Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer* **89**(2), 232–238.

Zhang, J., Baran, J., Cros, A., et al. (2011). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database* **2011**, bar026.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**(5), 365-376.

Smith, R., Brooks, D., Clegg, J., et al. (2015). The untapped potential of routine cancer data: A call for comprehensive research. *Journal of Cancer Research and Clinical Oncology* **141**(8), 1403-1412.

Goldstein, B.A., Navar, A.M., & Carter, R.E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European Heart Journal* **38**(23), 1805-1814.

Hoadley, K.A., Yau, C., Hinoue, T., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**(2), 291-304.e6.