A Survival Analysis of Character Mortality in Fictional Narratives: A Competing Risks Analysis

Project by:

Prashanthi R S

MSc Statistics

Abstract

This mini-project investigates the application of survival analysis techniques on a structured fictional dataset derived from the *Game of Thrones* character database, demonstrating the versatility of time-to-event models beyond clinical domains. The study utilizes Kaplan-Meier estimation to visualize baseline survival functions and applies log-rank tests to assess differences in survival across subgroups defined by attributes such as gender, nobility status, and house allegiance. To quantify the effect of these covariates on mortality risk, both univariate and multivariate Cox proportional hazards regression models are employed. Recognizing the complexity of character deaths within the narrative, the analysis is extended through competing risk models using cumulative incidence functions (CIF), offering more accurate estimates of cause-specific mortality while accounting for the interdependence of different risk events. This dual approach of survival and competing risk analysis provides a robust framework that mirrors methodologies commonly used in healthcare and epidemiological research.

All analyses were conducted using R software. By employing a fictional dataset, this project offers an accessible yet rigorous introduction to advanced statistical methods, serving as a practical foundation for future applications in more complex and sensitive real-world datasets.

List of Figures

Table of Contents

1. INTRODUCTION

Survival analysis is a statistical approach used to study the time until the occurrence of an event, often employed in medical and epidemiological research to investigate outcomes such as death, disease recurrence, or recovery. These methods not only help model and predict the timing of events but also uncover the factors influencing them. While survival analysis is traditionally rooted in clinical settings, its applications extend well beyond, offering powerful insights across a range of creative and unconventional domains.
This mini project explores one such unique application: analysing character mortality in the fictional world of the television series *Game of Thrones*. By leveraging mortality data from the show, this study demonstrates how survival and competing risk models can be applied to narrative-driven data. The show's intricate plotlines and detailed character arcs provide an engaging and accessible platform for applying advanced statistical techniques. Through this focused and compact project, the goal is to reinforce the conceptual understanding of survival analysis while showcasing its versatility beyond traditional research contexts.

1.1. Problem Description

This study seeks to explore the application of multiple survival analysis techniques—namely, Kaplan-Meier estimators, log-rank tests, Cox proportional hazards models, and Fine-Gray competing risk models—within a fictional but well-structured dataset. By examining various character attributes such as gender, house affiliation, allegiance, and nobility status, we aim to understand their influence on mortality risk within the Game of Thrones narrative.

The project also investigates the presence of competing risks, where different causes of death are mutually exclusive and may influence the time-to-event distribution. Through this lens, we model both cause-specific hazards and sub-distribution hazards, which are essential for providing accurate and interpretable risk estimates in multi-outcome environments. This analysis mirrors the real-world complexity seen in clinical or public health data, where multiple risk pathways coexist.

1.2. Project Scope

- Applying survival analysis methods to a non-clinical dataset.
- Implementing both standard survival models and competing risks models using statistical software tools such as R.
- Interpreting mortality patterns and risk factors within the context of a fictional universe.
- Drawing parallels between fictional and real-world datasets to understand the broader applicability of survival modeling techniques.

1.3. Dataset description

The dataset used in this project is publicly available on Figshare and contains detailed information about *Game of Thrones* characters across **Seasons 1 to 8**, which originally aired from **April 2011 to May 2019** on HBO.

- **Demographic and Social Information:**
  - **name**: Name of the character.
  - **sex**: Gender of the character (e.g., Male/Female).
  - **religion**: The religion or belief system followed by the character.
  - **occupation**: The character's role or job.
  - **social_status**: Social class or nobility status.
  - **allegiance_last**: Final allegiance or house the character is associated with.
  - **allegiance_switched**: Indicates if the character switched allegiances during the series.

- **Timeline and Event Data:**
  - **intro_season / intro_episode:** Season and episode in which the character is first introduced.
  - **intro_time_sec / intro_time_hrs**: Time of appearance in seconds and hours from the start.
  - **dth_season / dth_episode:** Season and episode in which the character died.
  - **dth_time_sec / dth_time_hrs:** Time of death in seconds and hours.
  - **censor_time_sec / censor_time_hrs**: Cumulative net running time at censoring or death of character in seconds/hours.
  - **exp_season / exp_episode**: Last season and episode in which the character was seen alive.
  - **exp_time_sec / exp_time_hrs**: Survival time of character in seconds/hours (until death or last seen).
  - **featured_episode_count**: Number of episodes in which the character appeared.
  - **prominence:** A measure of how central or prominent the character is in the storyline.
  - **dth_flag**: Binary indicator of whether the character died.

- **Death and Cause Related Details:**
  - **dth_description:** Narrative description of the character’s death.
  - **icd10_dx_code / icd10_dx_text**: ICD-10 diagnostic code and description
  - **icd10_cause_code / icd10_cause_text:** ICD-10 cause-of-death code and description.
  - **icd10_place_code / icd10_place_text**: ICD-10 location of death code and description.
  - **top_location / geo_location:** Geographic or regional location of the character’s death.
  - **time_of_day:** Approximate time of day when the event occurred (e.g., day/night, where applicable).

1.4. Objectives

- To explore the application of survival analysis techniques using a fictional yet structured dataset based on Game of Thrones character mortality.

- To estimate survival probabilities using Kaplan-Meier curves and compare survival across groups (e.g., gender, nobility status, house allegiance) using log-rank tests.

- To assess the impact of character traits on mortality using univariate and multivariate Cox proportional hazards models.

- To model cause-specific mortality using Fine-Gray sub-distribution hazard models for competing risks.

- To demonstrate the real-world relevance of statistical methods by extending survival and competing risks modeling beyond medical data into narrative-based, complex datasets.

2. LITERATURE REVIEW

[1] In the article "Understanding Survival Analysis: Kaplan-Meier Estimate," Goel, Khanna, and Kishore explain the Kaplan-Meier estimate as a valuable tool for survival analysis in clinical trials, especially for quantifying time to meaningful events, like death or disease relapse. They discuss the problems generally encountered in survival studies, i.e., the issue of censored observations, wherein subjects withdraw early from the study or do not get to experience the event before study closure. The authors refer to the capability of the Kaplan-Meier approach to make the best use of available data in such a case, allowing comparison of survival between treatment arms using statistical tests like the log-rank test. The authors also refer to the relevance of this method in Ayurvedic

research, wherein it allows one to test the effectiveness of different treatment interventions using survival data of the patients. Their research highlights the relevance of the Kaplan-Meier estimate to facilitating evidence-based practice in clinical and research settings.

[2] The paper "Survival Analysis Part I: Basic Concepts and First Analyses" by Clark et al. (2003) serves as an introductory guide to survival analysis, particularly in the context of cancer studies. It outlines the significance of survival time, which can refer to the duration from diagnosis to death or relapse, highlighting that many patients may not experience the event of interest by the end of a study, leading to challenges such as censoring. The text emphasizes the use of specialized statistical methods due to the non-normally distributed nature of survival data, which typically exhibit a skewed distribution of early and late events. Key methodologies discussed include Kaplan-Meier survival curves for estimating survival probabilities, log-rank tests for comparing survival distributions between groups, and Cox proportional hazards regression for assessing the influence of covariates on survival. Overall, the authors underscore the need for these advanced analytical techniques to accurately interpret survival data and derive meaningful conclusions in cancer research.

[3] Moolgavkar et al. (2018) in their paper "An Assessment of the Cox Proportional Hazards Regression Model for Epidemiologic Studies" critically evaluate the application of the Cox proportional hazards (PH) model in epidemiological research, focusing on its fundamental assumptions and the potential biases that arise when these assumptions are violated. They illustrate that the model's reliance on summary measures of exposure, such as pack-years of smoking, can obscure the true impact of temporal factors like intensity, duration, and time since cessation, thereby leading to misleading risk estimates. The authors emphasize that inadequate control for strong time-dependent confounding, such as that presented by cigarette smoking, can significantly bias relative risk (RR) estimates for modest correlated risk factors. As a solution, they advocate for using parametric

models that directly estimate hazard functions, thus allowing for a more nuanced understanding of time-varying exposure dynamics and improving the accuracy of risk assessments in epidemiological studies.

[4] Wolbers et al. (2014) in their paper "Competing risks analyses: objectives and approaches" provide a comprehensive overview of competing risks methodologies, emphasizing their importance in analysing the time to first observed events when multiple outcomes are possible. They highlight the need for appropriate statistical methods, such as the cumulative incidence function for descriptive statistics and various regression models, to obtain valid analyses in the presence of competing events. The paper underlines that traditional survival analysis techniques, like the Kaplan-Meier method, may overestimate risks when competing events are not addressed, advocating for regression approaches that effectively capture the impact of covariates on both the cumulative incidence and cause-specific hazards, thereby enhancing the understanding of clinical outcomes in cardiovascular studies and composite endpoints.

[5] In the realm of survival analysis, particularly in health research, the use of competing risks models has gained recognition for effectively handling scenarios where multiple events can prevent the occurrence of the primary event of interest. Fufa et al. in their study "Competing risk models to evaluate the factors for time to loss to follow-up among tuberculosis patients at Ambo General Hospital" (2023) employed both cause-specific hazard (CSH) and sub-distribution hazard (SDH) models to analyse the time until patients with tuberculosis were lost to follow-up, with death treated as a competing risk event. This approach allowed for a nuanced understanding of how various covariates—such as sex, residence, HIV status, and age—affect the time until patients discontinue treatment, highlighting the limitations of traditional survival analysis methods in the presence of competing risks.

[6] In the study "Factors associated with death in patients with tuberculosis in Brazil: Competing risks analysis" by Viana et al. (2020), a retrospective, population-based cohort study was conducted using data from 2008 to 2013 in Brazil to analyse factors associated with mortality among TB patients. The Fine & Gray sub-distribution model, based on the cumulative incidence function (CIF), was employed to identify factors associated with mortality among TB cases, while considering the probability of an event occurring before a specific time and maintaining observations on competing events within the range of risks. This model addresses the limitations of classical survival analysis by accounting for competing events, making it suitable for analysing factors associated with death in TB cases where individuals are at risk of multiple causes of death. Factors strongly associated with probable TB deaths included male gender, age over 60 years, illiterate schooling, black and brown colour/race, being from the Southern region, having mixed clinical forms, and alcoholism. HIV positive serology was also strongly associated with probable TB deaths. The study highlights the need for targeted surveillance and early case detection to reduce mortality among TB patients, leading to more timely detection and treatment.

[7] In "Survival analysis in the presence of competing risks," Zhongheng Zhang discusses methodologies for survival analysis when competing risks are present, which pose challenges because the hazard function lacks a direct link to the cumulative incidence function (CIF). The paper examines the use of the Fine-Gray model, which directly models the covariate effect on CIF, reporting sub-distribution hazard ratios (SHR), and compares it to the Cox proportional hazard model, which explores covariate effects on cause-specific hazards. It also highlights the limitations of the Kaplan-Meier estimator in the context of competing risks, as it overestimates the incidence of the event of interest. The article emphasizes appropriate modelling techniques and the use of R code for accurate statistical analysis, utilizing functions from the cmprsk and riskRegression packages.

[8] This article, titled "An Overview on the Complement of Kaplan-Meir Estimation and Cumulative Incidence Estimation in the Presence of Competing Risks Simulation Approach" by Valarmathi Srinivasan, Babu C Lakshmanan, and Chinnaiyan Ponnuraja, addresses methodological issues in clinical trials related to competing risks and compares the complement of Kaplan-Meier (1-KM) method with cumulative incidence estimation. The study uses simulated data with three competing events to demonstrate the differences between the two methods. It advocates for the use of cumulative incidence methods in clinical research to avoid overestimation of event probabilities, which can result from the misapplication of the Kaplan-Meier approach. The authors highlight the inappropriateness of using 1-KM to estimate failure probabilities in the presence of competing risks, supporting their argument with simulations and references to existing literature. They also use a real dataset of bone marrow transplant patients from Klein and Moeschberger (1997) to estimate the probability of disease progression using both 1-KM and cumulative incidence (CI). The paper concludes that in a competing risk setting, the complement of the Kaplan-Meier method overestimates the true failure probability, while cumulative incidence provides a more appropriate estimate

[9] The study by Gatechompol et al. (2022) employed the Cumulative Incidence Function (CIF) to estimate the probability of TB development after Antiretroviral Therapy (ART) initiation among PLWH in Thailand. Competing risks regression was used to calculate the subdistribution hazard for developing TB, with death considered as a competing risk. Risk time was defined as beginning 3 months after ART initiation to align with the exclusion criteria for unmasking TB occurrence, and it ended on the last visit date, the date of TB diagnosis, death or June 1, 2021, whichever came first. The subdistribution hazard models were applied to assess covariate effects on the CIF. This approach allowed the researchers to account for the fact that some participants might die before developing TB, which would prevent them from being counted as TB cases, thus providing a more accurate estimate of TB incidence in this population.

[10] The study by Shen et al. employed a sophisticated competing risk analysis methodology to investigate prognosis in head and neck basaloid squamous cell carcinoma (HNBSCC). Using data from the SEER-18 registry on 1,163 cases diagnosed between 2004-2013, the researchers implemented cumulative incidence function (CIF) methodology, which treats death from other causes as competing risks rather than censored events, thereby providing unbiased probability estimates of cause-specific mortality. This approach was enhanced through the application of Fine and Gray's proportional sub-distribution hazard model, enabling the development of a nomogram for individualized prediction of cause-specific mortality, with model performance rigorously evaluated using concordance indexes and calibration curves, and internal validation conducted through bootstrap resampling.

3. METHODOLOGY

The analysis involves a descriptive exploration of the dataset followed by the application of survival analysis techniques including the survival function, hazard function, Kaplan-Meier estimator, log-rank test, Cox proportional hazards model (both with independent and dependent variables), and competing risk analysis using the Fine and Gray model. These techniques help in estimating survival probabilities, assessing covariate effects, and understanding the impact of competing events on event occurrence.

3.1. Terminologies:

- **Survival Function:**

  The **survival function**, denoted as **S(t)**, represents the probability that a subject or individual will **survive beyond a specified time** point *t*. It is a central component of survival analysis, offering insight into the time duration until one or more events of interest (such as death, failure, or relapse) occur. The function essentially captures the distribution of survival times in a population.

  $$S(t) = P(T>t) = 1-F(t)$$

Where,

- T is the random variable representing survival time,
- F(t) is the **cumulative distribution function (CDF)**, representing the probability that the event has occurred by time *t*.
- Thus, S(t) expresses the complement of the CDF — i.e., the chance that the event **has not** occurred by time *t*.

  This function decreases over time and helps summarize the survival experience of a population, forming the basis for methods such as the Kaplan-Meier estimator and Cox proportional hazards model.

- **Hazard Function:**

  The **hazard function**, denoted as **h(t)**, represents the **instantaneous risk** of experiencing the event of interest (such as death, failure, or relapse) at a specific time *t*, **given that the individual has survived up to that time**. It provides the event rate at time *t* per unit of time, conditional on survival until that moment.

  $$h(t) = f(t) / S(t)$$

  S(t) is the survival function, f(t) is the distribution function of survival times.

  The hazard function is especially important in comparing the risks over time between groups and forms the basis of models such as the **Cox proportional hazards model**. It captures how the risk of event occurrence evolves, allowing for interpretations like constant, increasing, or decreasing hazards over time.

- **Kaplan Meier Survival Function:**

  The Kaplan-Meier (KM) estimator is a non-parametric method used to estimate the survival function from time-to-event data, especially when the data includes right-censored observations (i.e., subjects whose event time is unknown beyond a certain point). It is widely used in medical and reliability studies to understand the proportion of individuals surviving over time.

The KM estimator is calculated at each time an event occurs and updates the survival probability accordingly. The estimate remains constant between event times, producing a stepwise survival curve.

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

t_i : A specific time point at which one or more events occurred.

n_i : The number of individuals at risk just before time t_i.

d_i: The number of events that occurred at time t_i.

This estimator enables the visualization and comparison of survival experiences across different groups and is often paired with the **log-rank test** for hypothesis testing.

- **Log Rank Test:**

  The **log-rank test** is a **non-parametric hypothesis test** used to compare the **survival distributions** of **two or more groups**. It evaluates whether there is a statistically significant difference in the **survival experience** (i.e., time-to-event) between groups over the observed period.

  It specifically tests the **null hypothesis** that there is **no difference in the survival functions** across the groups — i.e., the survival probabilities at any time point are equal among the groups.

  The log-rank test is widely used because it accommodates **censored data** and does not assume any specific distribution for survival times, making it ideal for preliminary group comparisons before modelling (e.g., with the Cox model).

- **Cox Proportional Hazards Model (Time Independent variables):**
  The **Cox proportional hazards (PH) model** is a widely used **semi-parametric regression model** in survival analysis. It describes how one or more **independent covariates** affect the **hazard function** without requiring a specific baseline hazard distribution. The model assumes that the hazard for an individual is a product of a **baseline hazard** and a function of the covariates.
- The model estimates how each covariate **proportionally changes the hazard** of experiencing the event, holding other variables constant. It is particularly powerful because it handles **right-censored data** and allows for the inclusion of continuous and categorical predictors.

$$h(t|X) = h_0(t) * exp(\beta^T X)$$

If $\exp(\beta_i) > 1$, the covariate **increases** the hazard (i.e., risk).

If $\exp(\beta_i) < 1$, the covariate **decreases** the hazard.

❖ **h(t|X)**:
The **hazard function** at time *t* for an individual with covariate vector X. It gives the **instantaneous risk** of the event occurring at time *t*, given that the individual has survived up to time *t*.

❖ $\boldsymbol{h_0(t)}$:
The **baseline hazard function** — the hazard function when all covariates are zero. It is unspecified (non-parametric) and represents the underlying risk over time without the influence of covariates.

❖ **β**:
A **column vector of regression coefficients** (parameters) associated with the covariates. Each βi represents the **log hazard ratio** corresponding to the covariate X.

❖ **X**:
A **column vector of covariates** (independent variables), which could include demographic, clinical, or behavioural factors.

The model is called **proportional** because the hazard ratios between individuals remain constant over time, assuming the **proportional hazards assumption** holds.

- **Cox Proportional Hazards Model (Time Dependent variables):**

  When covariates are time-dependent, the Cox proportional hazards model is extended to incorporate variables whose values change over time. This extension enables the modelling of more realistic scenarios where an individual's risk factors (covariates) may vary during the follow-up period. For example, changes in medication, disease status, or behavioural patterns.

  $$h(t|X(t)) = h_0(t) * exp(\beta^T X(t))$$

  This model allows for covariate values to be **updated continuously or periodically**, making it suitable for dynamic risk environments.

  It accommodates **internal covariates** (functions of survival time) and **external covariates** (not directly influenced by the survival process).

    - h(t|X(t)) : Hazard at time *t* given covariates that vary over time.
    - $h_0(t)$: Baseline hazard function.
    - X(t) : Covariate vector **as function of time**.
    - $(\beta^T X(t))$: Linear predictor with time-varying covariates.

- **Proportional Hazards Assumption:**

  The Proportional Hazards (PH) assumption is a key premise of the Cox proportional hazards model. It states that the hazard ratio between any two individuals remains constant over time, regardless of the absolute value of the hazard at any point in time. In other words, the relative effect of the covariates on the hazard function is time-invariant.

  - The covariates have a multiplicative effect on the hazard function.
  - The baseline hazard function can change over time, but the ratio of hazards between individuals does not.
  - Violation of this assumption suggests the need for alternative modeling techniques, such as stratified Cox models or time-varying covariates.

- **Competing risks analysis:**

  In **competing risks** analysis, an individual is at risk of experiencing **one of several mutually exclusive events**, where the **occurrence of one type of event precludes the occurrence of other types**. For example, in medical studies, a patient may be at risk of dying from multiple causes (e.g., cardiovascular disease or cancer), but only one cause of death can be observed.

  When competing risks are present, using standard survival methods (like Kaplan-Meier or Cox regression assuming a single event) can lead to **biased or inflated estimates of survival probabilities**, because they treat competing events as censored rather than as distinct outcomes.

  This type of analysis is essential when:

  - **Multiple failure types** (events) are possible.
  - **Accurate estimation of cause-specific or sub-distribution risks** is required.
  - **Clinical or policy decisions** depend on understanding the probability of each event type.

- **Cumulative Incidence Function (CIF):**

  The Cumulative Incidence Function (CIF) estimates the probability of experiencing a specific type of event by time t in the presence of competing risks**.** Unlike the standard survival function, which considers only a single type of event, the CIF accounts for the fact that other competing events may occur and prevent the event of interest from happening.

  Mathematically, the CIF for cause k at time t, denoted $F_k(t)$, is:

  $$F_k(t) \, l= P\ (T \leq t, \text{event}=k)$$

  CIF gives the cumulative probability that an individual will experience the event of type k by time t, considering the presence of other competing events.

- **Fine and Gray Model**

  The **Fine and Gray sub-distribution hazard model** is an extension of the Cox proportional hazards model designed specifically for **competing risks data**. Unlike cause-specific hazard models that treat competing events as censored, the Fine and Gray model directly models the **sub-distribution hazard**, which allows for **correct estimation of the cumulative incidence function (CIF)** for a specific cause in the presence of competing events.

  - Provides **direct modelling** of CIF.
  - Allows **proper interpretation** of covariate effects on the probability of the event of interest.
  - Especially suitable for **clinical decision-making**, where **real-world probabilities** of event occurrence are important.

- **Sub-distribution Hazard:**

  The sub-distribution hazard function, denoted as $\lambda_j(t|X)$ for cause j given covariates X, models the instantaneous risk of failure due to cause j at time t within a pseudo-population of individuals who have not yet experienced failure from cause j. This pseudo-population includes those who are still event-free and those who have experienced a competing event (and are thus considered to be "still at risk" of the event of interest in this specific model). Consequently, the sub-distribution hazard does **not** adhere to the traditional definition of a hazard rate based on those truly at risk. Instead, it directly models the cumulative incidence function (CIF) for the event of interest.

- **Interpretation of Sub-distribution Hazard (sHR):**

  The sub-distribution hazard ratio (sHR), derived from the Fine and Gray model, quantifies how a unit change in a covariate multiplies the sub-distribution hazard for a specific event of interest, relative to a baseline. Unlike a standard hazard ratio, the sHR's interpretation must always be considered within the context of competing risks and its direct link to the Cumulative Incidence Function (CIF). An sHR greater than 1 suggests that higher values of the covariate are associated with a faster cumulative incidence of the event of interest, even in the presence of other failure types. Conversely, an sHR less than 1 indicates a slower cumulative incidence. Importantly, the sHR reflects the covariate's influence on the *probability* of experiencing the specific event by a given time, acknowledging that individuals might experience other competing events instead. Therefore, its impact is on the overall trajectory of the CIF for that particular cause.

3.2. Application to Dataset

- **Survival and hazard functions** will be used to estimate the probability of survival and the instantaneous event rate over time for the studied population.
- The **Kaplan-Meier estimator** will calculate non-parametric survival curves, allowing visualization and comparison across groups.
- **Log-rank tests** will evaluate whether survival distributions differ significantly between groups (e.g., treatment or demographic categories).
- **Cox proportional hazards models** (both independent and time-dependent) will assess the impact of covariates on event occurrence while checking the **proportional hazards assumption**.
- **Competing risks analysis**, including **CIFs**, and the **Fine-Gray model**, will be applied to account for multiple event types and to model the probability of failure from a specific cause.

3.3. Data Preparation and Analysis

In the initial stage of data preparation, the dataset underwent a thorough restructuring. **Categorical variables** were **re-coded and grouped** based on relevant **ICD-10 cause categories**, ensuring that each entry could be accurately associated with the clinical condition or event it represented. This categorization helped in simplifying the interpretation of results and ensuring consistency in data analysis. For the **competing risk analysis**, a dedicated variable was created to represent different types of death event. Any **missing or blank entries** in the death event variable were systematically replaced with the label **"Alive"** to represent censored observations. This approach was critical in ensuring that the data remained consistent and appropriately handled for **survival analysis**. Moreover, by ensuring a clear distinction between event types, this data preparation step laid the groundwork for the **Cox proportional hazards model** and **Fine and Gray model** where **exp_time_hrs** was used exclusively as the **time-to-event variable**. Given the use of these **semi-parametric models**, no assumption regarding the parametric distribution of exp_time_hrs was required.

For the **Cox proportional hazards model**, the **proportional hazards assumption** was verified using **Schoenfeld residuals**, ensuring the model's validity. The **Fine and Gray model** was applied to account for **competing risks** and to estimate the **sub-distribution hazard ratio** for different causes of death.

3.4. Software used

- **Microsoft Excel**: Used for initial data cleaning, such as removing missing or inconsistent entries, formatting columns, renaming variable labels for clarity, and preparing the dataset structure before importing it into R for analysis.

- **R (survival, survminer, cmprsk, dplyr, ggplot2)**: Used for conducting survival and competing risk analysis. This included Kaplan-Meier survival curves, log-rank tests, univariate and multivariate Cox proportional hazards models, time-dependent Cox models, Cumulative Incidence Functions (CIFs), and Fine-Gray models. Visualization of results was done using ggsurvplot() and CIF plots.

4. RESULTS AND DISCUSSION

The **Results and Discussion** section will present a comprehensive analysis comprising descriptive statistics to summarize the key characteristics of the dataset, followed by Kaplan-Meier survival curves and log-rank tests to compare survival distributions across groups. Additionally, Cox proportional hazards regression will be employed to identify significant predictors of survival time. To account for competing events, a competing risk analysis will also be conducted, offering deeper insights into cause-specific outcomes.

4.1 Descriptive Statistics

This section provides a summary of the demographic and clinical characteristics of the study population. Variables such as age, gender, event status, and cause of death were analysed. For continuous variables like exposure time (in hours), measures such as mean, median, and standard deviation were calculated. Categorical variables were summarized using frequencies and percentages.

4.1.1 Interpretation of Categorical Variables

Sex Distribution:

Out of the total 359 individuals, 254 (70.8%) were male and 105 (29.2%) were female. No unknown values were recorded for sex.

Social Status:

Of the 359 characters analysed, 247 (68.8%) were classified as Lowborn while 112 (31.2%) were Highborn, indicating a predominance of lower social status in the cohort.

Religion:

A diverse set of religious affiliations were recorded. The Faith of the Seven was the most common (35 individuals), followed by the Old Gods (31), Great Stallion (19), Lord of Light and Drowned God (14 each). A large portion (233) were recorded as "Unknown," which may influence the interpretation.

Occupation:

The dataset included individuals identified as "Silk Collar" (96), "Boiled Leather Collar"

(221), and 42 marked as "Unknown". Boiled Leather Collar, possibly indicating working-class or military status, made up the majority.

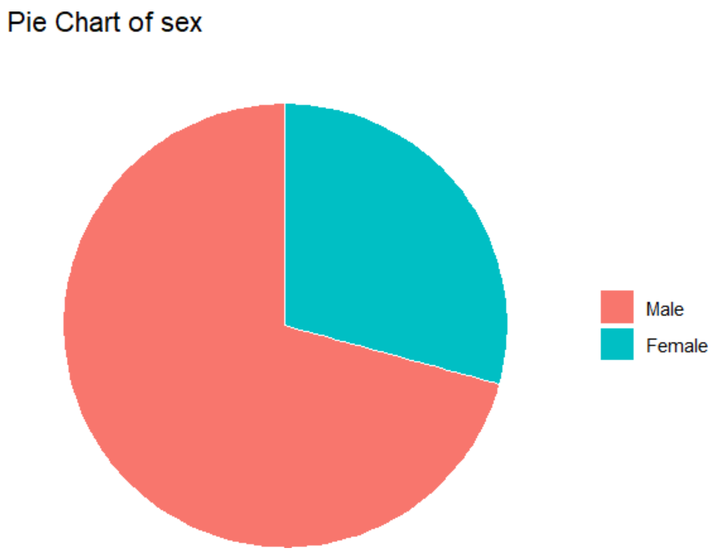

Figure 1: Pie Chart representing distribution of sex

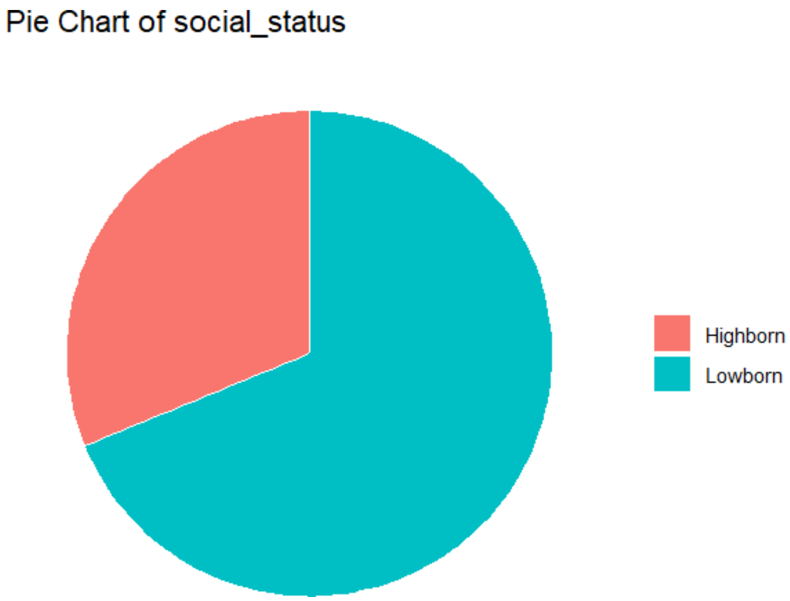

Figure 2: Pie Chart representing distribution of social status

Cause of Death (Cause Category):

Among the deceased, the most frequent cause was assault-related (119), followed by war-related (61), and legal execution (11). Self-harm and other external causes were comparatively less frequent. 147 individuals were still alive.

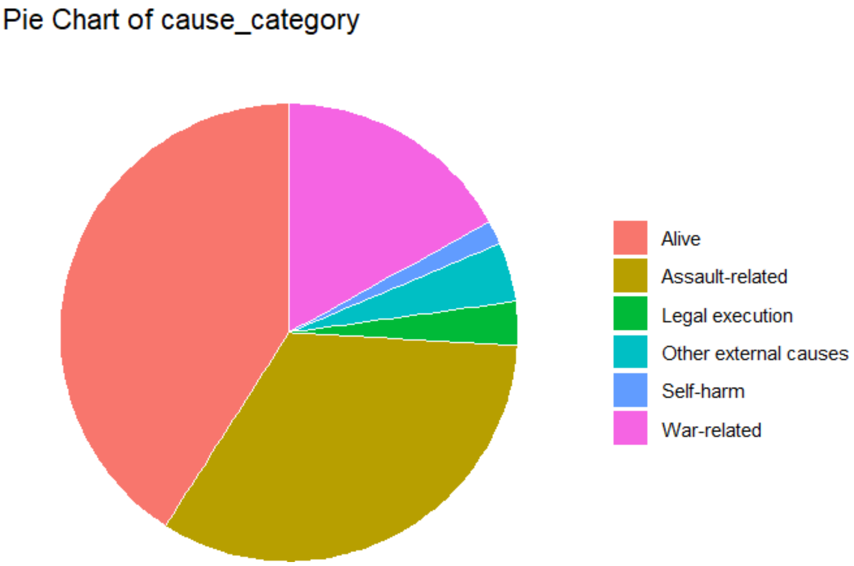

Figure 3: Pie Chart representing distribution of causes of death

Last Allegiance:

The most frequent final allegiance was "Other" (135), followed by House Stark (60), Lannister (34), and Night's Watch (27). Only a small number were aligned with Targaryen, Frey, Greyjoy, and Bolton. The allegiance of 37 individuals remained unknown.

Allegiance Switched:

Majority (304) of the individuals remained loyal to a single house or group, while 55 had switched allegiance at least once.

Event Status:

At the end of the study period, 212 individuals (59.1%) were dead and 147 (40.9%) were alive.

4.1.2. Interpretation of Continuous variables

❖ Censor Time (censor_time_hrs)

- Values range from 0.10 to 64.11 hours.
- Median = 52.91: Half of the censored individuals were observed up to ~53 hours.
- Mean = 46.92: Indicates overall moderate censoring times.
- Most subjects who were censored survived quite long in the study.

❖ Exposure Time (exp_time_hrs)

- Values range from 0.00 to 63.99 hours.
- Median = 18.49, Mean = 22.68.
- Right-skewed distribution: a few had very long exposure durations.
- Majority were exposed for short-to-moderate durations before event or censoring.

❖ Prominence

- Values range from 0.11 to 7.34.
- Median = 0.88, Mean = 1.13.
- Right-skewed again: a small number of highly prominent individuals.
- Most individuals had low prominence, suggesting limited storyline focus.

❖ Featured Episode Count

- Ranges from 1 to 67 episodes.
- Median = 3, Mean = 7.81.
- Strongly skewed right: few characters appeared in many episodes.
- The dataset includes both minor and major characters; majority were featured briefly.

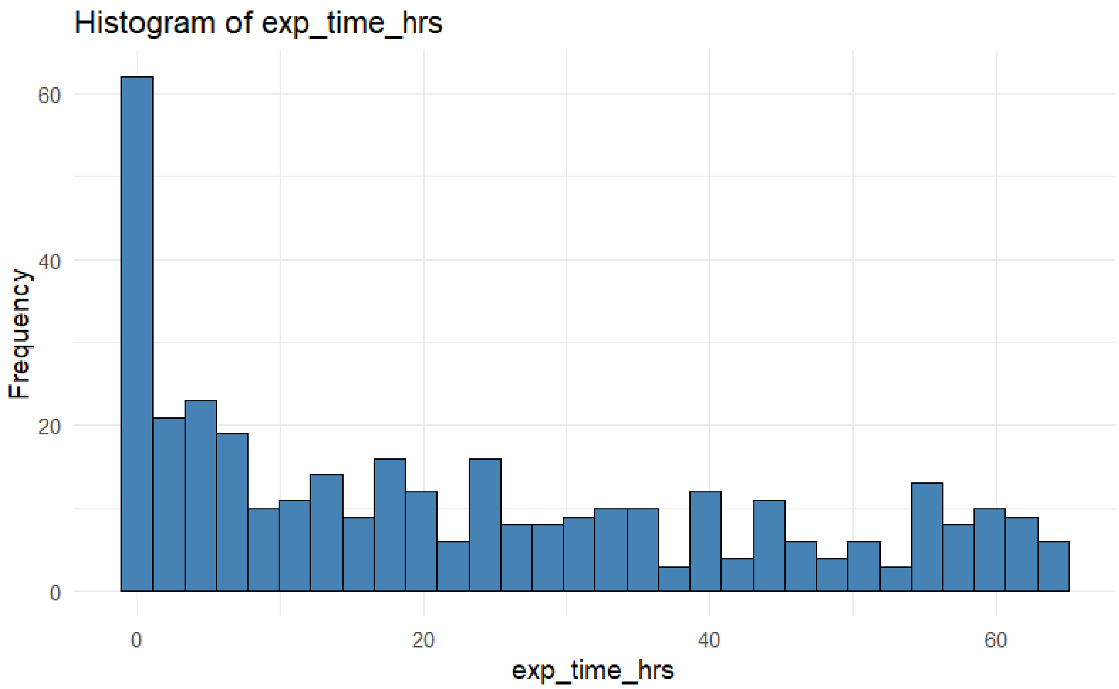

Figure 4:Histogram of survival time of characters in hours

This histogram displays exposure time data with a strong right-skewed distribution. The majority of observations occur within the first 5 hours (over 60 cases), followed by a long tail extending to about 65 hours.

4.2 Survival Analysis Using Multiple Approaches

To evaluate time-to-event outcomes in the dataset, multiple survival analysis methods were applied. The **Kaplan-Meier estimator** was used to estimate survival probabilities, and the **Log-Rank test** compared differences between groups. The **Cox proportional hazards model** identified significant predictors of hazard over time. Additionally, the **Fine-Gray model** was employed to handle competing risks affecting the event of interest.

4.2.1 Kaplan Meier Curves and Log rank tests for various variables

- Kaplan Meier Survival Curve for the entire cohort for the variable exp_time_hrs

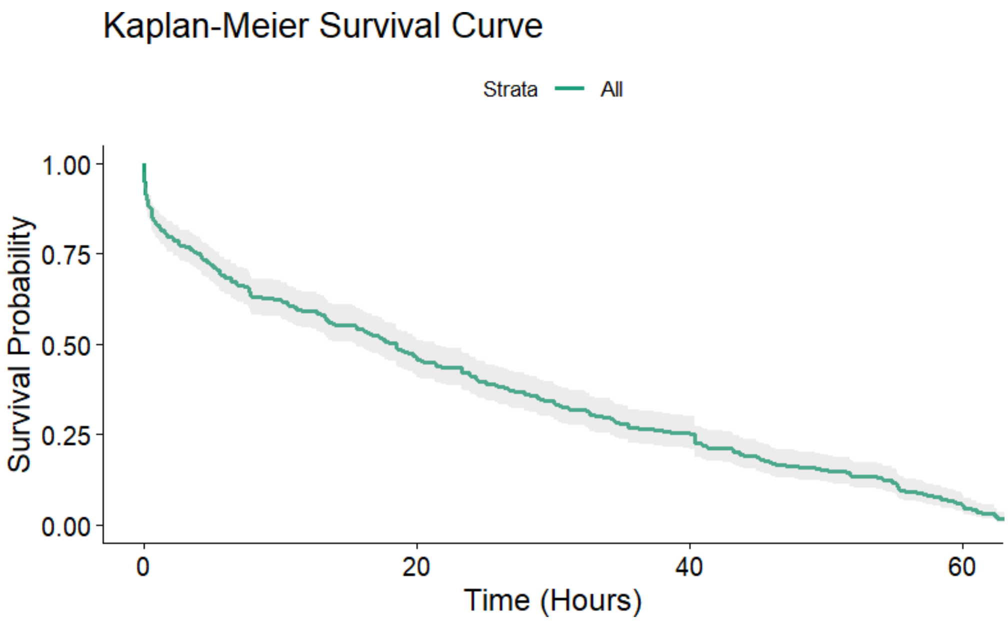

Figure 5: KM curve for the cohort

For the above graph, survival probability steadily declines from nearly 100% at time 0 to about 0% at 60 hours. The median survival time (50% probability) occurs at approximately 20 hours. The gray shaded area represents the confidence interval, which widens slightly over time.

- **KAPLAN MEIER SURVIVAL PROBABILITY STRATIFIED BASED ON SEX**

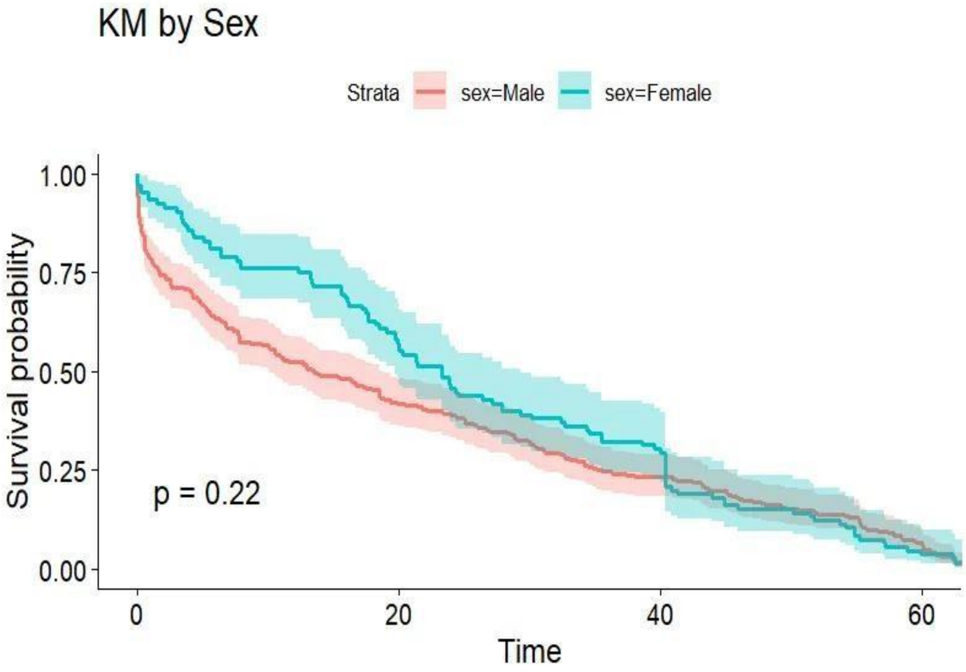

Figure 6: Stratified KM by sex

Females (blue) show higher survival rates than males (red) until about 40 hours, when curves converge. Median survival is approximately 15 hours for males versus 25 hours for females. Log rank test (p=0.22) indicates failure to reject the null hypothesis, suggesting no statistically significant difference in survival between males and females despite visual differences in the curves.

- **KAPLAN MEIER SURVIVAL PROBABILITY STRATIFIED BASED ON THE GROUPS OF INDIVIDUALS WHO SWITCHED ALLEGIANCE DURING THE SHOW**

Individuals who switched (Group 2 (blue)) demonstrates dramatically better survival than who didn't switch (Group 1 (red)). Median survival for Group 2 is approximately 45 hours versus only 10 hours for Group 1. Log rank test (p<0.0001) strongly rejects the null hypothesis, confirming allegiance switching is a highly significant predictor of survival with Group 2 having substantially lower risk.

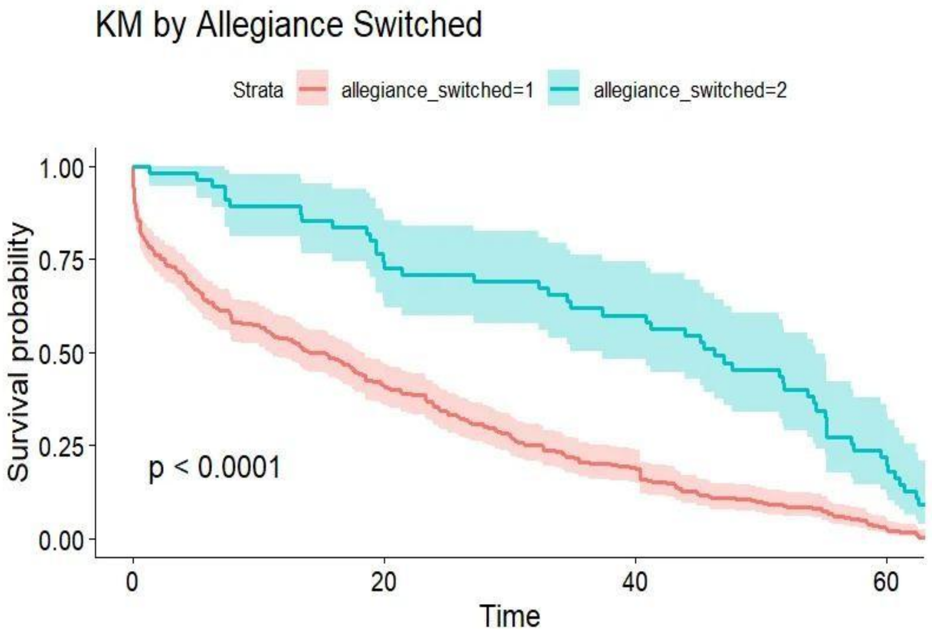

Figure 7: KM Stratified by Allegiance switched

- **KAPLAN MEIER SURVIVAL PROBABILITY STRATIFIED BASED ON THE GROUPS BASED ON SOCIAL STATUS**

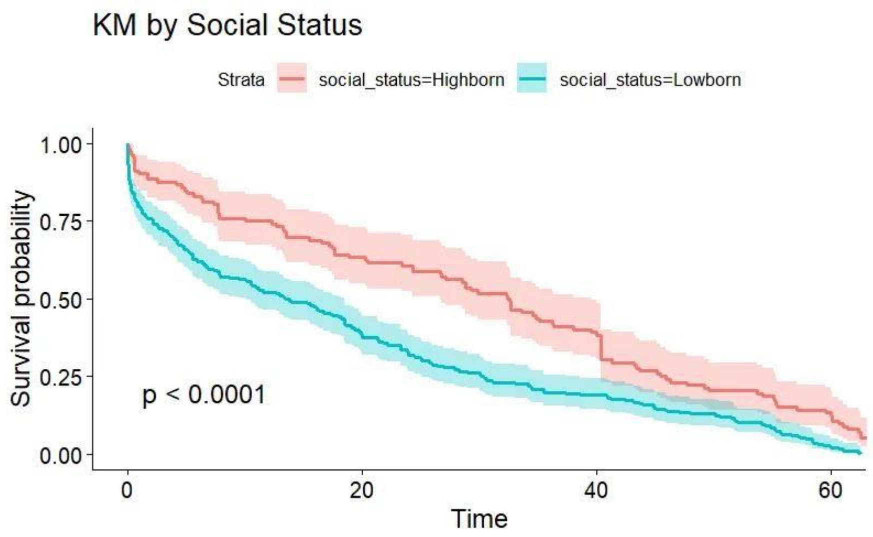

Figure 8: KM stratified by social status

Highborn individuals (red) consistently show better survival than Lowborn (blue). Median survival is approximately 35 hours for Highborn versus 15 hours for Lowborn. Log rank test (p<0.0001) strongly rejects the null hypothesis, confirming social status significantly impacts survival with Highborn individuals having substantially lower risk than Lowborn.

4.2.2 Cox Proportional Hazard Model

- **UNIVARIATE COX PROPORTIONAL MODEL:**

To explore the relationship between sex and the survival outcome, a univariate Cox proportional hazards model was fitted.

Model Overview

- Number of observations (n): 359
- Number of events (deaths): 212

  - Predictor variable: sex

- The hazard ratio (HR) for sex is 0.534, with a 95% confidence interval of 0.383 to 0.745.
- The p-value is 0.00022, indicating that the variable is highly statistically significant ($p < 0.001$).

Interpretation:

- The HR of 0.534 implies that males have approximately 46.6% lower hazard (risk of death) compared to females, assuming the coding is such that males are the reference group (you may want to confirm your reference coding).
- The confidence interval does not cross 1, which further confirms statistical significance.
- The result indicates a protective effect of male gender on survival in this dataset. Model Diagnostics
- Concordance = 0.571: This indicates the model has a moderate discriminatory ability, better than random but not highly predictive.
- All three global tests (Likelihood Ratio, Wald, and Score/Log-rank) show strong statistical significance, confirming the variable's relevance in predicting survival outcomes:
  - **Likelihood Ratio Test**: $\chi^2 = 15.24$, $p < 0.001$
  - **Wald Test**: $\chi^2 = 13.62$, $p < 0.001$
  - **Score Test**: $\chi^2 = 14.07$, $p < 0.001$

The proportional hazards assumption was assessed using Schoenfeld residuals. The p-value for the variable sex (p = 0.072) and the global test (p = 0.072) indicated no significant violation of the assumption, suggesting that the proportionality of hazards over time holds for the model.

- **MULTIVARIATE COX PROPORTIONAL MODEL**:

A Cox proportional hazards regression was performed to assess the effect of sex, occupation, and allegiance_last on the hazard of the event (death) over time among 359 individuals, with 212 events observed.

Overall Model Fit:

- The Likelihood Ratio Test ($\chi^2 = 19.8$, df = 3, p = 0.0002), Wald Test (p = 0.0006), and Score Test (p = 0.0004) all indicate that the model is statistically significant.
- The model's concordance is 0.607, suggesting a moderate ability to distinguish between high-risk and low-risk individuals.

Interpretation of Covariates:

- Sex:
  - Coefficient: -0.556
  - Hazard Ratio (HR): 0.573 (95% CI: 0.4009 – 0.8197)
  - p-value: 0.0023 (Significant)
  - Interpretation: Females have a 43% lower hazard of death compared to the reference group males), adjusting for occupation and allegiance.
- Occupation:
  - Coefficient: -0.060
  - HR: 0.941 (95% CI: 0.8689 – 1.0200)
  - p-value: 0.140 (Not significant)
  - Interpretation: Occupation does not have a statistically significant impact on the hazard of death in this model.
- Allegiance_Last:
  - Coefficient: 0.0366
  - HR: 1.037 (95% CI: 0.9901 – 1.0867)
  - p-value: 0.1237 (Not significant)

  - Interpretation: A one-unit increase in the allegiance score is associated with a 3.7% increase in hazard, but this effect is not statistically significant.

- ❖ The PH assumption was tested, individually, none of the covariates violate the PH assumption (all p > 0.05).
- ❖ The global test is just barely significant at p = 0.046, indicating a mild overall violation, but not strong enough to discard the model.

- **TIME DEPENDENT COX PROPORTIONAL MODEL:**

- Considering the fact that certain events or covariates change over time, the allegiance_switched variable captures such a time-varying effect by adjusting the hazard rate dynamically based on whether the character has switched allegiance during the course of the show.

- Characters who switch allegiance may experience different risks at various points in the show. For example, a character who was initially loyal to one group but later switches may face different challenges or risks during that time. This dependency on time is crucial because a simple static (non-time-varying) variable wouldn't account for the changes in risk that arise from switching allegiances.

- The **interaction term** (tt(allegiance_switched)) is used to model how the effect of switching allegiance changes as time progresses. In the results, tt(allegiance_switched) has a significant p-value **(0.000319)**, indicating that switching allegiance significantly affects the hazard rate over time.

Model Interpretation of Hazard Ratios:

- Sex:
    - For example, the hazard ratio for sex Female is 0.8187, meaning that, controlling for other variables, females have a lower hazard (lower risk of death) compared to males.
- Religion:
    - For religion Old Gods, the hazard ratio is 0.4442, indicating that characters following the Old Gods have a lower risk of death compared to those following Great Stallion.
- Occupation:
    - For occupation Boiled leather collar, the hazard ratio is 0.8181, suggesting that characters with this occupation have a slightly lower risk of death compared to those with the Silk collar occupation.
- Social Status:
    - The hazard ratio for social_status Lowborn is 1.7453, indicating that Lowborn characters have a higher risk of death compared to Highborn characters.
- Allegiance Last:
    - For allegiance_last Greyjoy, the hazard ratio is 2.7484, indicating that characters aligned with Greyjoy have a higher risk of death compared to Stark characters.
- Allegiance Switched:
    - The hazard ratio for tt(allegiance_switched) is 0.8376, which shows that characters who switched allegiance experience a lower risk of death over time, suggesting that allegiance switching may have a protective effect or alter the individual's survival dynamics.
    - The time-varying effect of allegiance_switched is significant signifying its effect on the hazard changes as time progresses.

4.2.3 Competing Risks Model

- **CUMULATIVE INCIDENCE FUNCTION:**

The competing risks analysis examines the survival patterns of characters in this fictional universe using a competing risks framework, specifically through Cumulative Incidence Functions (CIF).

The dataset categorizes character deaths into five distinct causes:

1. **Assault-related**: Direct violence from other characters (e.g., sword fights, assassinations)
2. **War-related**: Deaths occurring during formal military conflicts
3. **Other external causes**: Torture, animal attacks, exposure to elements
4. **Legal execution**: Formal executions ordered by authority figures
5. **Self-harm**: Suicide or self-inflicted wounds

Unlike traditional survival analysis, competing risks methodology acknowledges that characters face multiple possible causes of death, each "competing" with the others to determine a character's fate. The CIF presented here shows the probability of experiencing each specific type of death over time, accounting for the presence of these competing risks.

Interpretation of the Cumulative Incidence Function

The CIF graph reveals several key patterns in character mortality.

Assault-related deaths (red line)

This category represents the dominant cause of mortality, reaching approximately 45% cumulative incidence by 60 hours. The steepest increase occurs within the first 10 hours, reflecting the brutal and sudden nature of interpersonal violence in the series. This pattern aligns with the show's reputation for unexpected character deaths through assassinations, ambushes, and one-on-one combat.

War-related deaths (blue dashed line)

War represents the second most common cause of death, with cumulative incidence reaching about 25% by 60 hours. The curve shows a notable acceleration after the 50-hour mark, potentially corresponding to major battles featured later in the series (e.g., Battle of the Bastards, Battle of Winterfell). This demonstrates how large-scale conflicts become increasingly significant mortality drivers as the narrative progresses.

Other causes of death

The remaining three categories—other external causes (green dotted), legal execution (purple dash-dot), and self-harm (orange)—each contribute minimally to overall mortality, remaining below 5% throughout the observation period. This suggests that while these death types feature memorably in the narrative, they represent relatively rare events compared to direct violence and warfare.

This mortality pattern reflects the series' central themes: the dangerous nature of power struggles, the devastating impact of warfare, and the constant threat of betrayal. The analysis quantifies what fans have long observed—in the game of thrones, death comes swiftly and often through violence rather than natural causes.

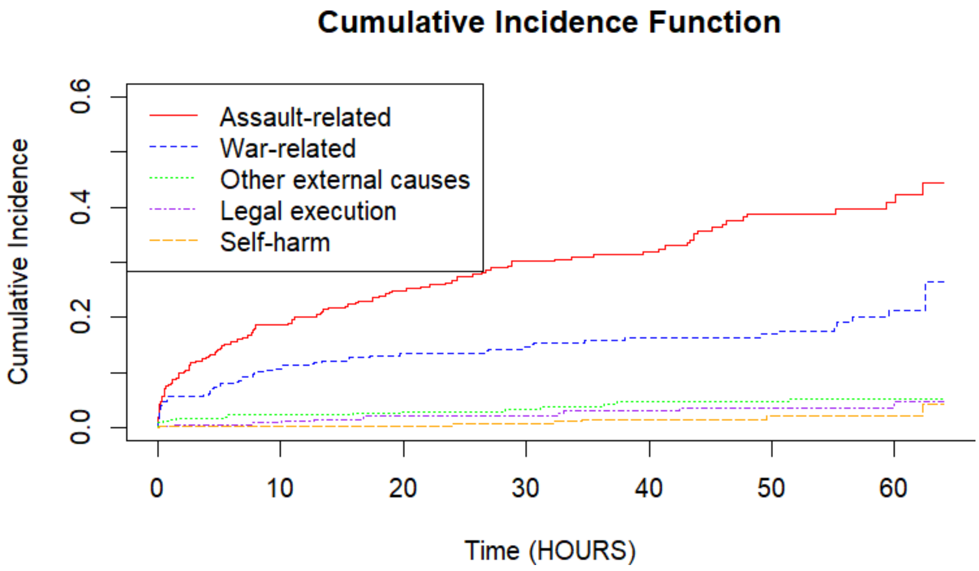

Figure 9: CIF for different causes of death

- **FINE – GREY SUB-DISTRIBUTION HAZARD MODEL:**
  - **Overview:**

    This study applied the Fine-Gray sub-distribution hazard model to assess the impact of individual-level factors—namely, sex, allegiance-switching status, and social status—on the probability of cause-specific mortality in a historical population. The analysis was conducted in the context of competing risks, where each individual was at risk of experiencing one among several possible causes of death.

    Although multiple causes of death were present in the dataset, this report focuses specifically on two causes:

  - **Assault-related deaths**
  - **War-related deaths**

    The Fine-Gray model was chosen because it accounts for the presence of competing events, producing sub-distribution hazard ratios (SHRs) that estimate how covariates affect the cumulative incidence of a particular event type over time.

  - Reasons for proceeding with univariate analysis first:
    1. Understand their individual associations with each event type.
    2. Identify which variables merited inclusion in the final multivariate model.
    3. Avoid unnecessary multicollinearity or overfitting.

    Based on theory and initial results, all three predictors—sex, allegiance-switching, and social status—were retained for the multivariate analysis.

  - Interpretation of Sub-distribution Hazard Ratios (SHRs)

A sub-distribution hazard ratio (SHR) greater than 1 indicates a higher cumulative incidence of the specific cause of death over time, while a SHR less than 1 suggests a lower cumulative incidence, i.e., a protective effect.

Event 1: Assault-Related Deaths

Univariate Model Results:

- Sex (male vs. female): SHR = 0.49, 95% CI: (0.31, 0.76), p = 0.002.Males had 51% lower sub-distribution hazard of dying from assault than females.
- Allegiance-switching (yes vs. no): SHR = 0.44, 95% CI: (0.24, 0.80), p = 0.008.Those who switched allegiances had a 56% reduction in the cumulative incidence of assault-related death.
- Social status: SHR = 1.21, 95% CI: (1.01, 1.45), p = 0.04.Individuals with higher social status had a 21% increased risk of dying from assault.

Multivariate Model Results:

- Sex (male): SHR = 0.53, 95% CI: (0.34, 0.84), p = 0.007.Even after adjusting for other covariates, males were still 47% less likely to die from assault.
- Allegiance-switching: SHR = 0.47, 95% CI: (0.25, 0.86), p = 0.02.The protective effect of allegiance-switching persisted, with a 53% reduction in assault-related deaths.
- Social status: SHR = 1.14, 95% CI: (0.94, 1.39), p = 0.18.Social status was no longer a statistically significant predictor, suggesting its effect may have been confounded by sex or allegiance.

Interpretation:

Being male and having switched allegiance were both independently associated with a significantly lower risk of dying from assault, even when adjusting for social status. The effect of social status observed in univariate analysis did not hold in the multivariate context, indicating that higher social status alone does not significantly predict assault mortality when sex and allegiance are taken into account.

Event 2: War-Related Deaths

Univariate Model Results

- Sex (male): SHR = 0.69, 95% CI: (0.45, 1.07), p = 0.10.Males were 31% less likely to die from war-related causes, but the association was not statistically significant.
- Allegiance-switching: SHR = 0.67, 95% CI: (0.35, 1.30), p = 0.24. Those who switched allegiances had 33% lower hazard, but again, this was not statistically significant.
- Social status: SHR = 1.06, 95% CI: (0.87, 1.30), p = 0.56.No strong or significant effect of social status was observed.

Multivariate Model Results

- Sex (male): SHR = 0.72, 95% CI: (0.47, 1.12), p = 0.14
- Allegiance-switching: SHR = 0.73, 95% CI: (0.37, 1.43), p = 0.36
- Social status: SHR = 1.02, 95% CI: (0.83, 1.26), p = 0.84

None of the covariates in the multivariate model were statistically significant predictors of war-related deaths.

Interpretation:

Although the direction of effects was consistent (e.g., males and allegiance-switchers had reduced risk), none of these relationships reached statistical significance. This may suggest that individual-level factors were less predictive of war-related deaths, which may have been influenced more by contextual or external factors (e.g., geographic proximity to battle, collective decisions, military strategy) not captured in the model.

5. CONCLUSION

This comprehensive survival analysis of Game of Thrones characters reveals that mortality in the series is not just a function of time but deeply influenced by nuanced character traits and narrative dynamics. The Kaplan-Meier estimates show a consistent decline in survival probabilities, while stratified analyses highlight that females and Highborn characters enjoy a statistically significant survival advantage. The inclusion of time-dependent covariates in the Cox model underscores the importance of shifting allegiances—characters who adapt and realign their loyalties improve their chances of survival over time, reflecting the series' emphasis on political astuteness. Interestingly, static variables such as occupation and original allegiance do not significantly affect survival, suggesting that static affiliations matter less than strategic adaptability. The Fine-Gray competing risks model further refines our understanding by identifying assault and war as the leading causes of death, and shows how factors like social status and allegiance switching alter the probabilities of dying from specific causes. Collectively, these results paint a vivid statistical portrait of a fictional world where survival hinges on a complex interplay of gender, birthright, and one's ability to navigate an ever-shifting political landscape—aptly capturing the essence of the Game of Thrones narrative.

Limitations:

- **Unknown Data**: The use of "9" to represent unknown values in key variables, such as cause of death, could lead to misinterpretation and may have impacted the analysis, as it may not fully capture the complexities of the unknown category.
- **Sparse Categories in Cause of Death**: The cause_category variable had very few cases in certain categories (e.g., "legal - execution" and "self-harm"), which made the cumulative incidence functions (CIFs) difficult to interpret and led to potential biases in hazard estimation for these rare causes.

REFERENCES

[1] P. K. J. K. Manish Kumar Goel, “Understanding Survival Analysis: Kaplan-Meier Estimate,” *International Journal of Ayurveda Research,* vol. 1, no. 4, pp. 274-278, 2010.

[2] M. B. S. L. a. D. A. TG Clark, “Survival Analysis Part I: Basic concepts and first analyses,” *British Journal of Cancer,* vol. 89, no. 2, pp. 232-238, 2003.

[3] E. T. C. H. N. W. a. E. C. L. Suresh H. Moolgavkar, “An Assessment of the Cox Proportional Hazards Regression Model for Epidemiologic Studies,” *Risk Analysis,* vol. 38, no. 4, pp. 779-790, 2018.

[4] M. T. K. V. S. S. B. S. K. J. J. K. L. G. H. Marcel Wolbers, “Competing risks analyses: objectives and approaches,” *European Heart Journal,* vol. 35, no. 42, p. 2936–2941, 2014.

[5] F. D. B. a. T. A. Diriba, “Competing risk models to evaluate the factors for time to loss to follow-up among tuberculosis patients at Ambo General Hospital,” *Archives of Public Health,* vol. 81, pp. 81-117, 2023.

[6] N. S. P. D. A. M. V. L. S. B. A. L. d. S. B. P. C. B. Paulo Victor de Sousa Viana, “Factors associated with death in patients with tuberculosis in Brazil: Competing risks analysis,” *PLOS ONE,* vol. 15, no. 10, 2020.

[7] Z. Zhang, “Survival analysis in the presence of competing risks,” *Annals of Translational Medicine,* vol. 5, no. 3, 2017.

[8] B. C. L. C. P. Valarmathi Srinivasan, “An Overview on the Complement of Kaplan-Meir Estimation and Cumulative Incidence Estimation in the Presence of Competing Risks_Simulation Approach,” *International Scientific Research Journal,* vol. 1, no. 6, pp. 61-65, 2015.

[9] J. S. S. U. A. A. F. v. L. F. C. a. S. J. K. Sivaporn Gatechompol, “Incidence and factors associated with active tuberculosis among people living with HIV after long-term antiretroviral therapy in Thailand: a competing risk model,” *BMC Infectious Diseases,* vol. 22, no. 346, 2022.

[10] N. S. L. Y. Weidong Shen, “Cause-specific mortality prediction model for patients with basaloid squamous cell carcinomas of the head and neck: a competing risk analysis,” *Journal of Cancer,* vol. 9, no. 21, pp. 4009-4017, 2018.

[11] Kleinbaum, David G. and Mitchel Klein. Survival Analysis: A Self-Learning Text. Third Edition. New York: Springer, 2012.