# Identifying Injury Risk Factors for Elite Soccer Teams Using Survival Analysis

Anna Linnea Jarmann


Thesis submitted for the degree of
Master in Informatics: Programming and System Architecture
60 credits

Department of Informatics
The Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

Spring 2023

# Identifying Injury Risk Factors for Elite Soccer Teams Using Survival Analysis

Anna Linnea Jarmann

Identifying Injury Risk Factors for Elite Soccer Teams Using Survival Analysis

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

# Abstract

Soccer is a sport millions of people enjoy worldwide, with a significant amount of resources devoted to injury prevention for player health and team performance. Various methods are used to assess injury risk as part of injury prevention. One commonly employed method is using the training load metric Acute Chronic Workload Ratio (ACWR) to measure injury risk. However, its limitations have sparked discussions and scepticism. Machine learning techniques have also emerged as a method of injury prevention by recognising injury risk factors and predicting injuries. However, few studies have explored the use of survival analysis for this task. This thesis aims to fill this gap by extending survival analysis beyond its traditional use in medical research to injury risk assessment in sports. We investigate injury risk factors in subjective training load and wellness data from two elite female soccer teams and extract the variables with the most significant impact on injury outcomes. We experiment with different approaches and apply the Cox Proportional Hazards Model (CPH) to estimate the magnitude of each variable's impact on injury risk. We also explore time-varying analysis using the Cox Time-Varying Model (CTV). Our results show that combining recurrent injuries with averaged variables from all days prior to injury provides the most accurate and reliable results, also allowing for time-varying analysis. We perform feature selection using regularisation to extract the most significant factors for injury risk, which include prior injuries, sleep quality, fatigue, and ACWR. Using cross-validation, we determine the optimal penalty term for regularisation based on the lowest Bayesian Information Criterion (BIC) and highest Concordance Index (C-index). Our research significantly contributes to computer and sports science by offering a novel approach for extracting injury risk factors in a dataset using survival analysis. We deliver valuable insights into the factors affecting injury risk in female soccer players. Additionally, our results provide possibilities for developing targeted injury prevention programs and improving player health and performance in various sports and injury types.

# Acknowledgments

I would like to thank my supervisors, Matthias Boeker, Cise Midoglu, Pål Halvorsen and Steven Hicks, for their support, guidance and encouragement. Their valuable feedback, ideas and advice have truly shaped my work. A special thank you goes to Matthias for introducing me to survival analysis which has made this research truly interesting and enlightening.

I also want to thank my family for their constant love and encouragement that has helped me stay motivated throughout my degree.

Finally, I would like to thank my dear kollektiv for all the support and for always turning tough times into good times.

# Contents

# List of Figures

# List of Tables

# Acronyms

**ACWR** Acute Chronic Workload Ratio.

**BIC** Bayesian Information Criterion.

**C-index** Concordance Index.

**CPH** Cox Proportional Hazards Model.

**CTV** Cox Time-Varying Model.

**FFRC** Female Football Research Centre.

**sRPE** Session Rating of Perceived Exertion.

# Chapter 1

# Introduction

## 1.1 Motivation

Soccer is a globally popular sport that demands a significant amount of resources and attracts a massive fan base [15]. The sport's popularity has led to the involvement of not only players, trainers and physicians but also researchers, investors, football fans and bettors.

The optimal physical condition of the players is essential to the sport's success, and coaches and medical staff play a critical role in ensuring this. One of the important factors in maintaining performance during matches and training is minimising the risk of injury [31]. Preventing injuries is necessary for athletes' health and the opportunity to have longevity in their careers. Injuries can have long-term effects on an athlete's physical and mental health, leading to chronic pain, decreased mobility, and possibly ending their careers too early. Therefore, keeping athletes injury-free or reducing the severity of their injuries is critical to their success. Injuries are also costly, and rehabilitating soccer players to return to their previous level of play demands significant resources.

Women's soccer has yet to achieve the same level of attention as men's, and research on female athletes is scarce [9]. Female athletes are more susceptible to specific injuries, such as ACL (Anterior Cruciate Ligament) tears, a common knee injury in sports with sudden movements [27]. Studying female soccer players' injury patterns is critical to developing targeted injury prevention programs that can improve their performance and career longevity. The lack of investment in women's soccer has also contributed to inadequate training facilities, medical staff, and equipment, increasing injury risk. Therefore, investing in women's soccer and providing the same resources and support as male soccer can enhance female athletes' success and promote injury prevention.

In recent years, injury prediction in soccer has gained significant attention among researchers, and several methods have been employed to identify potential risk factors and develop predictive models. Machine learning has emerged as a promising approach that can analyse large amounts of data to identify patterns and predict injuries effectively [60]. Another common approach for evaluating the risk of injury is using the

Acute Chronic Workload Ratio (ACWR), which compares an athlete's current fatigue to their fitness levels [25]. While ACWR is widely used, its accuracy in predicting injuries has been questioned in some sports, including soccer [63].

One promising approach for evaluating injury risk is survival analysis. Survival analysis is a statistical method used to analyse time-to-event data, such as the time until an injury occurs [48]. It estimates the probability of an event occurring at a particular time, considering the censoring of data when it has not occurred by the end of the study. Survival analysis models can incorporate time-independent covariates, such as age and gender, and time-dependent covariates, such as training load, fatigue and stress levels, to predict the event's occurrence more accurately.

While machine learning has shown potential in identifying patterns and predicting injuries, survival analysis offers a complementary approach that can provide valuable insights into the risk factors and dynamics of injury occurrence in soccer. Despite its potential, there has been a lack of use of survival analysis for sports injuries, and more research is needed in this field [48]. Therefore, in this thesis, we aim to apply survival analysis techniques to identify injury risk factors for soccer teams.

## 1.2 Problem Statement

Based on the abovementioned issues, we want to explore using different survival analysis techniques to identify injury risk factors in soccer teams. The main research question we want to answer is:

> *How can we extract injury risk factors from training load and wellness data from elite female soccer teams using survival analysis?*

To do this, we aim to experiment with multivariate survival analysis models using training load and wellness metrics as covariates to identify which are most significant to injury outcome. We must first validate the dataset for survival analysis using univariate models. The dataset we use is from two female soccer teams with reported training load and wellness over the course of two years. We also want to use both first and recurrent injuries to compare these and find which yields the best results. Additionally, we want to compare the use of values from the day of the injury and averaged values from all days prior to the injury. As we use time-dependent covariates, we want to explore options to include the changes in these and how they may impact the injury outcome. When we find the best approach, we aim to perform feature selection using regularisation to extract the most significant injury risk factors.

To achieve our goal, we have formulated seven sub-questions that will help us to answer the main research question:

**RQ1.** How should missing data entries be handled?

**RQ2.** Can we use a univariate survival model to validate our dataset?

**RQ3.** Does the size of the dataset affect the results?

**RQ4.** Does the number of covariates used in the multivariate models affect the results?

**RQ5.** Should we use first injuries or recurrent injuries?

**RQ6.** Should we use covariate values from the day of the injury or from its whole duration interval?

**RQ7.** What penalty term and threshold results in optimal feature selection?

The goal is to apply computer science techniques to sports data to improve injury prevention strategies. By using survival analysis to analyse training load and wellness data, our research aims to identify the most significant risk factors for injuries in female soccer players. This can inform training and recovery programs and ultimately reduce injuries. Furthermore, by investigating questions related to data processing, missing data, dataset size, covariate selection, and feature selection using regularisation, this research aims to contribute to developing best practices in computer science. Overall, the goal is to leverage computer science techniques to improve athlete health and performance in soccer.

## 1.3 Scope

This thesis focuses on using survival analysis to identify injury risk factors in female elite soccer teams. The data used in this research is collected from two teams in the Norwegian elite soccer league, "Toppserien", from the years 2020 and 2021, and is obtained from the SoccerMon dataset [42]. This dataset includes information on player injuries and training load and wellness, which our analysis use as covariates. Our analysis identifies injury risk factors for teams as a whole rather than for individual players. We also investigate the impact of missing data on our results and explore feature selection and regularisation techniques. Our findings contribute to a better understanding of injury risk factors in elite female soccer teams and provide insights that can inform injury prevention strategies.

## 1.4 Research Methods

This thesis aims to identify injury risk factors by experimenting with survival analysis techniques. To accomplish this, we have adopted the research methodology of the ACM's (Association for Computing Machinery) teaching discipline for computer science. This methodology involves the development of a theoretical framework, designing and implementing a system, and evaluating the system's performance [17]. The ACM's discipline consists of three paradigms:

**Theory:** The first paradigm is rooted in mathematics and involves four steps for developing a valid theory: defining the objects of study,

hypothesising possible relationships among them, verifying them through proof, and interpreting results.

**Abstraction:** The second paradigm is rooted in the experimental scientific method and involves four stages for investigating a phenomenon: formulating a hypothesis, constructing a model and making predictions, designing an experiment and collecting data, and analysing results.

**Design:** The third paradigm, design, is rooted in engineering and involves four steps for constructing a system or device to solve a given problem: stating requirements, specifying them, designing and implementing the system, and testing it.

Our research uses the abstraction paradigm to experiment with multivariate survival analysis models using training load and wellness metrics as covariates to identify the most significant injury risk factors. We use the design paradigm to design and implement a system to identify injury risk factors and test its performance. By using the ACM methodology in our thesis, we aim to develop a systematic and rigorous approach to identifying injury risk factors in female elite soccer teams using survival analysis techniques.

## 1.5 Ethical Considerations

When collecting subjective training load and wellness data from soccer players, it is crucial to prioritise ethical considerations such as obtaining informed consent, ensuring transparency, and protecting privacy. The data used in this research is extracted from the SoccerMon dataset, which collects data from soccer players through pmSys [42]. Informed consent was obtained from each player, and they were fully informed about what data was being collected and how it will be used, ensuring transparency in the research process. To protect the privacy of the players, the collected data was anonymised and stored with unique anonymous IDs, and all personally identifiable information was removed from the dataset. The study behind the SoccerMon dataset has been certified by the Norwegian Privacy Data Protection Authority and is entirely anonymous, exempting it from further user consent. Moreover, the study does not include bio-bank, medical, or health data related to illness or interfering with the normal operation of the players, making it exempt from approval from the Regional Committee for Medical and Health Research Ethics - South East Norway and the Regional Committee for Medical and Health Research Ethics - Northern Norway. Since the data is anonymous, the dataset is publicly shareable based on Norwegian and General Data Protection Regulation (GDPR) laws. Overall, the ethical considerations were considered and appropriately handled, ensuring that the data collection and analysis adhered to the relevant regulations and ethical standards.

## 1.6 Main Contributions

In this thesis, we propose a method for identifying injury risk factors in soccer teams using survival analysis techniques. By analysing subjective data such as training load and wellness data, we aim to identify factors contributing to injury risk. Our method can benefit players, coaches, and medical staff by better understanding the factors affecting injury risk and enabling them to create effective injury prevention strategies. Our method applies to soccer and other sports where injury risk is a concern, making it a versatile technique for injury prevention.

Additionally, our research contributes to the field of computer science by providing alternative methods to evaluate risk and identify significant factors in a dataset. We also extend the application of survival analysis beyond its traditional use in medical research to a sports science context, a relatively new and growing area of research.

Furthermore, our research contributes to sports science by identifying injury risk factors for female elite soccer teams. While previous research has focused on injury risk factors in male soccer teams, there is limited research on female soccer teams. Our findings could fill this knowledge gap and improve injury prevention strategies for female soccer teams.

## 1.7 Thesis Outline

The rest of this thesis is organised into the following chapters.

**Chapter 2: Background and Related Work.** This chapter provides information on athlete monitoring and highlights the sources of our dataset, the Female Football Research Centre (FFRC), pmSys and Soccermon. Further, it discusses ACWR as an injury risk estimate. The chapter also presents related work using ACWR, machine learning and survival analysis.

**Chapter 3: Methodology.** This chapter presents the methods used in this thesis, including survival analysis techniques, univariate and multivariate models, data structures, techniques for feature selection and model evaluation, and tools used for implementation.

**Chapter 4: Experiments and Results.** This chapter presents and discusses our four experiments and their respective results using the methods presented in Chapter 3.

**Chapter 5: Discussion.** This chapter discusses insights from our findings, revisits the research questions, proposes potential use and addresses concerns and limitations with our approach. It also provides an overview of next steps and contributions.

**Chapter 6: Conclusion.** This chapter summarises and completes our thesis.

# Chapter 2

# Background and Related Work

Based on our motivation and aim for this thesis presented in the previous chapter, this chapter covers background information and related work that will support our methods and experiments. We do this by providing a comprehensive understanding of athlete monitoring and how the FFRC collects and utilises athlete monitoring data from elite female soccer players, including injury, training load, and wellness data. In this thesis, we highlight the pmSys system and the SoccerMon dataset, which is the data source for our experiments.

Moreover, we discuss current techniques and related work in injury prevention, focusing on using ACWR, machine learning, and survival analysis techniques. By reviewing the literature on injury prevention and related techniques, we can identify gaps in knowledge and develop novel approaches to injury prevention.

By presenting background information on the data we utilise and related work, this chapter sets the stage for the proposed methods of identifying injury risk factors using survival analysis techniques in elite female soccer teams.

## 2.1 Athlete Monitoring

Athlete monitoring is a process that involves the collection and analysis of data related to an athlete's physical and physiological performance to inform decisions related to training, recovery, and injury prevention [26]. This includes the collection of data such as training load, wellness, nutrition, and sleep. The data is then analysed to recognise patterns and trends that can be used to optimise performance, prevent injuries, and enhance recovery. Athlete monitoring is widely used in elite sports to help coaches, medical staff, and athletes make informed decisions about training, recovery, and injury prevention.

Two categories of data can be collected through athlete monitoring: subjective and objective [55]. Subjective data is provided by the athletes themselves, often through self-reported answers and evaluations of their health, wellness, and athletic performance. This includes data on how the athletes feel and perceive their performance. Objective data is collected

through physical measurements, such as tracking time, distance, and quantities, and provides direct data on the athlete's physical performance.

The data collected through athlete monitoring provides a vast amount of information and many possibilities. For instance, it can be used to improve player health and performance and prevent injuries. Additionally, the data can be used in research related to sports science, medicine, artificial intelligence, and women's health.

## 2.2 Female Football Research Centre

FFRC [19] is a research centre founded by UiT - The Arctic University of Norway with SimulaMet's Department of Holistic Systems as a research partner. The centre focuses on developing non-invasive and privacy-respecting technology for athlete monitoring to gain analytical insights from various perspectives, including biomechanics, sports-specific science, medicine, coaches, and athletes. FFRC aims to gain a new and fundamental understanding of factors that impact elite female soccer players' performance and overall health.

To collect monitoring data, FFRC uses pmSys, a system that collects subjective parameters like training load, wellness, injury, and illness [35][21]. Players complete a questionnaire via a mobile application, and the data is transmitted to a cloud-based backend system. This data is visualised in a trainer portal, where coaches and staff can view individual player and team data.

FFRC also gathers objective data, such as movement patterns and distances, using GPS metrics. These metrics are gathered using the STATSports APAX system, where players wear vests equipped with tracking devices during training sessions and matches. The system captures metrics such as acceleration distance, high-speed distance, and total distance.

FFRC aims to integrate female-specific parameters and introduce more automated analysis techniques like machine learning. This would enable researchers to gain deeper insights into the collected data and identify correlations and observations that could be used to enhance the performance of elite female soccer players. One of the automated analyses would be to predict future injuries.

## 2.3 PmSys

PmSys is developed through a collaboration between students and researchers at the Simula Research Laboratory, the University of Tromsø, and ForzaSys [35]. PmSys is a smartphone-based tool that enables systematic, long-term monitoring of athletes' physical and self-reported parameters. Using the system, players can report their wellness daily through a smartphone application, and coaches can access the team's data through a web-based trainer portal. The collected data allows coaches to understand the team's fitness and each player's progress. They can also identify which

players are peaking and which are not and adjust their training sessions and match strategies accordingly. Additionally, the collected data has been used to train a machine-learning model for predicting peak readiness in athletes [67].

### 2.3.1 Smartphone Application

The pmSys application is accessible to players through their smartphones, providing them with various reporting options. The application includes reports for corona checks, Session Rating of Perceived Exertion (sRPE), wellness, participation, injuries, illnesses, and game performance [35]. However, this thesis will only focus on the reports for sRPE, wellness, and injuries.

sRPE is a subjective measure of an individual's perceived level of exertion during a session of physical activity or exercise [22]. Individuals rate their perceived exertion on a scale of 0-10, with higher numbers indicating a higher level of exertion. In pmSys, players report their RPE for a specific session by answering questions about various factors, including the time since the session, session duration, type of session, and their perceived exertion. This information is used to calculate the overall sRPE for a given session, which provides an estimate of the training load and exertion for the player. Figure 2.1 shows how the reporting of sRPE looks in the application.

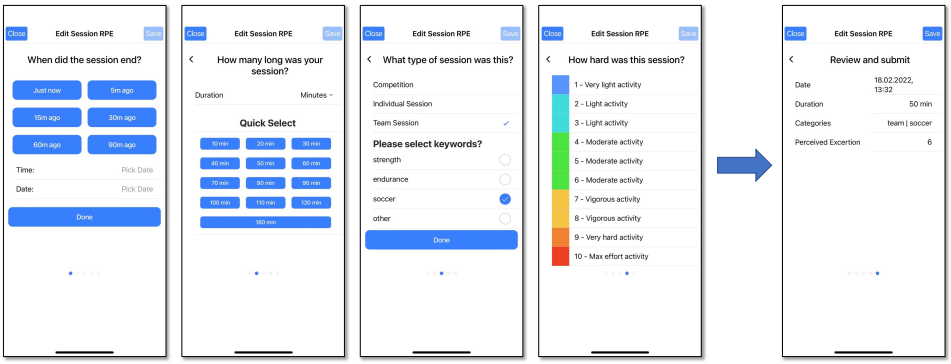



Figure 2.1: PmSys sRPE questionnaire [42]

For reporting wellness, the players complete a survey with questions regarding their readiness to play, fatigue, soreness, sleep quality, sleep duration, mood and stress levels, as shown in Figure 2.2. Readiness is reported on a scale of 1-10, with 1 indicating "not ready at all" and 10 indicating "can't wait!". Sleep duration is reported as the number of hours the player slept. Table 2.1 displays the rest of the wellness variables and what each reported value on a balanced Likert-scale of 1 to 5 indicates [42].

Players report injuries using a body silhouette feature available in the pmSys application. The feature allows them to pinpoint the location of their injury, as demonstrated in Figure 2.3. Players also choose whether it

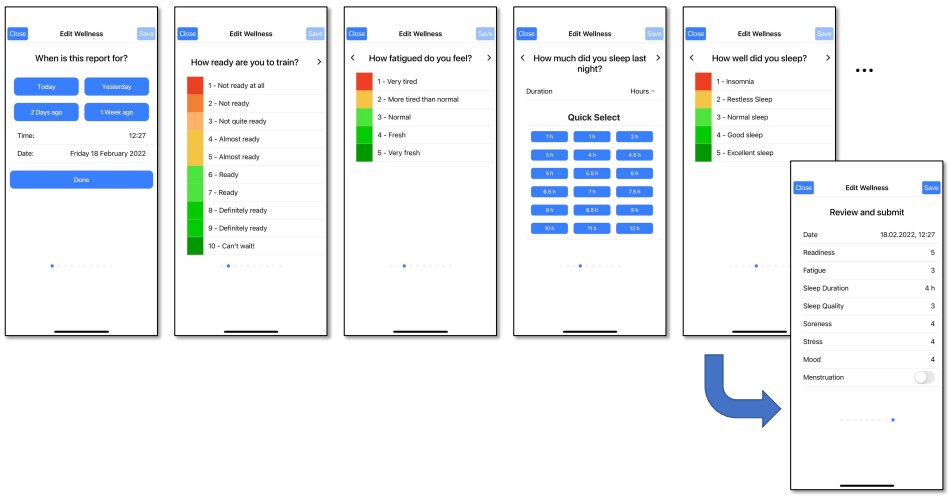



Figure 2.2: PmSys wellness questionnaire [42]

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Fatigue** | Very tired | More tired than normal | Normal | Fresh | Very fresh |
| **Soreness** | Very sore | A bit sore | Normal | Feeling good | Feeling great |
| **Sleep quality** | Insomnia | Restless sleep | Normal | Good | Very restful |
| **Mood** | Very bad mood | Bad mood | Normal | Good mood | Very good mood |
| **Stress** | Highly stressed | Somewhat stressed | Normal | Relaxed | Very relaxed |

Table 2.1: Scale of wellness variables [67]

was a minor or major injury and report the date and time of the injury, even if it occurred at a different time.

### 2.3.2 Web Based Trainer Portal

The trainer portal, a web-based Single-Page Application, allows team personnel, coaches, and physicians to observe the self-reported data collected from the players [21]. The portal displays the data for individual players or the entire team, including visualisations of injuries and illnesses and plots of key training load indicators, as illustrated in Figure 2.4.

## 2.4 SoccerMon

Researchers at the Simula Research Laboratory have developed a comprehensive collection of data called SoccerMon, which includes the physical and mental well-being and training load of soccer athletes [42]. The dataset includes subjective and objective measures, with subjective data collected

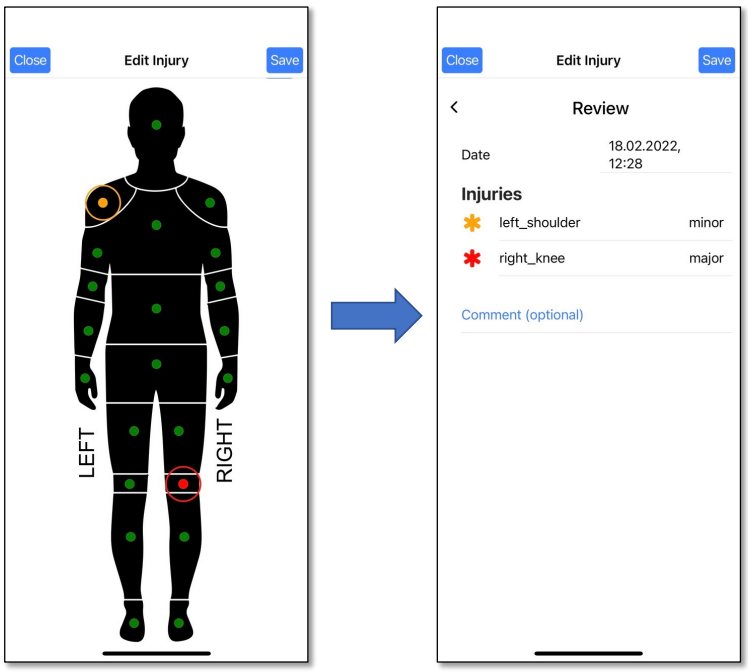



Figure 2.3: PmSys injury reporting [42]

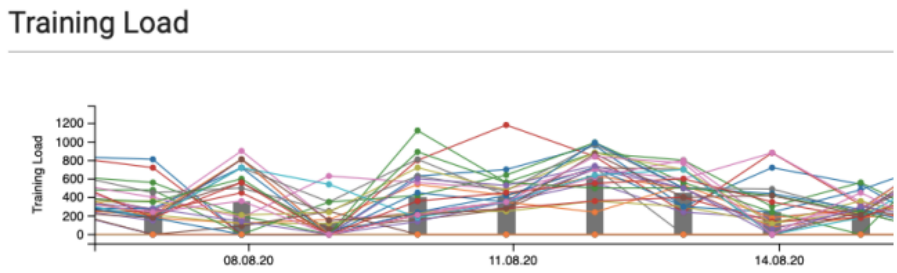
through pmSys and objective data through the STATSports APEX system.

SoccerMon is the largest dataset currently available that contains these types of data specifically for women's soccer, making it an invaluable resource for researchers and coaches. With little research done in this field for women's soccer, the dataset can be used to identify differences between men's and women's training responses and other health parameters, which can aid in developing individualised training plans and help prevent injuries.

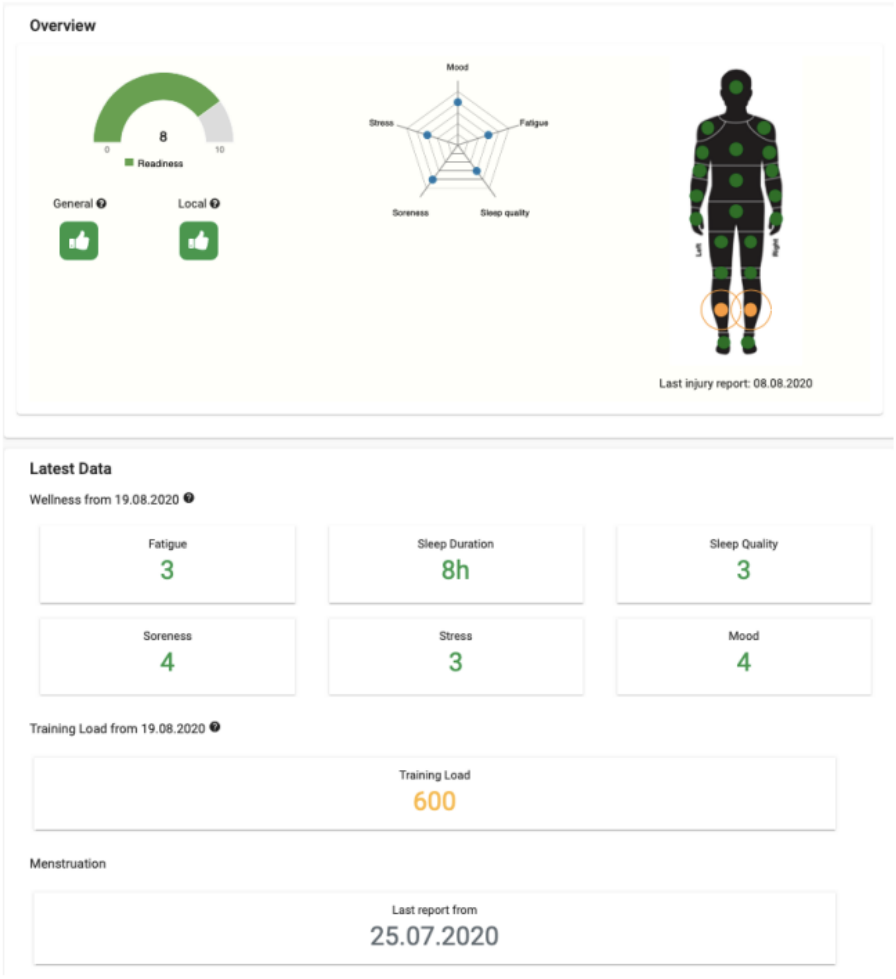
### 2.4.1 Subjective Data

The subjective data in the SoccerMon dataset is collected using pmSys, meaning it consists of daily reports of training load, wellness and injuries. The data is stored locally on the athletes' phones and then synced to a cloud server using the Amazon AWS public cloud [42]. To ensure privacy and compliance, each report is owned by the individual user and can only be accessed by coaches and staff through the trainer portal after the user attaches their data to a team. The data is then fully exported with athlete identification to match with objective data before the identification is replaced with random IDs to protect privacy.

The training load variables in the SoccerMon dataset are calculated from the player's sRPE. Table 2.2 shows the different variables in the training load data and how they are calculated.




**(a) Team overview**




**(b) Single player**

Figure 2.4: PmSys trainer portal (courtesy of ForzaSys AS)

### 2.4.2 Objective Data

The objective data for SoccerMon is collected using the STATSports APEX system, which tracks players' movements and positions through devices embedded in vests [42]. The device utilises multiple satellite systems,

| Training Session Reports | Description | Formula |
|---|---|---|
| Session RPE (sRPE) | The workload of a single session depending on the duration and the reported RPE values | $RPE \times duration$ |
| Training load (TL) | The sum of sRPE during a day | $\sum sRPE$ per day |
| Weekly Load (WL) | The sum of sRPE over the last 7 days | $\sum sRPE$ per week |
| Acute Training Load (ATL) | The current level of fatigue (average sRPE over the last 7 days) | $\frac{1}{7}\sum_{n=i}^{i+7} DL_i$ |
| Chronic Training Load (CTL) | The cumulative training dose that builds up over a longer period of time (average sRPE over the last 28 or 42 days) | $\frac{1}{x}\sum_{n=i}^{i+x} DL_i$ x = 28 or 42 |
| Acute Chronic Work Load (ACWR) | An indication of whether an athlete is in a well-prepared state, or at an increased risk of getting injured (ATL divided by CTL) | $\frac{ATL}{CTL}$ |
| Monotony | Reflection of training variation across the last 7 days (mean sRPE divided by the standard deviation (SD) = ATL / SD) | $\frac{ATL}{SD}$ |
| Strain | Reflection of the overall training stress from the last 7 days (total weekly sRPE multiplied with Monotony) | $WL \times Monotony$ |

Table 2.2: Training load metrics [42]

including GPS, GLONASS, Galileo and BeiDou, to gather data during training sessions and matches. After each session, the collected data is retrieved and uploaded to the club's laptop through the STATSports software. The data is then uploaded to OneDrive in a compressed format for the SoccerMon dataset.

### 2.4.3 Combining Subjective and Objective Data

The researchers further looked into combining subjective and objective parameters to give a more holistic view of the soccer players [42]. Creating a correlation matrix showed that the objective acceleration distance strongly correlated with the subjective metrics of readiness, sleep quality, and stress levels. Additionally, they observed a correlation between perceived exertion and GPS tracking metrics, with players feeling less exhausted with higher acceleration distance and more exhausted with increased sprints, high-speed distance, and total distance.

They note that this dataset has many potential research opportunities, including examining trends and correlations between various parameters [42]. They also suggest that further analysis of injury data with additional parameters could aid in injury prediction and prevention.

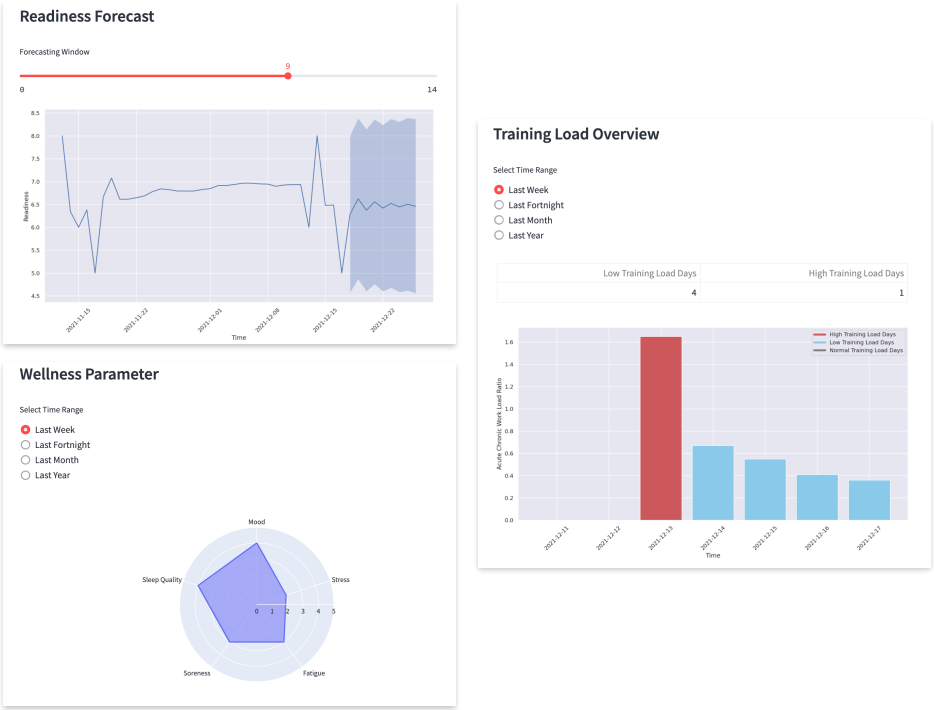



Figure 2.5: Soccer Dashboard player analyses and visualisations [10]

## 2.5 Soccer Dashboard

The SoccerMon dataset contains a large amount of data on female soccer players regarding wellness, health and training load. To understand and explore this data, researchers from Simula Metropolitan Center for Digital Engineering have developed Soccer Dashboard, a web-based, open-source tool for data visualisation and analysis [10]. Soccer Dashboard allows easy data aggregation, statistics generation, and advanced investigations into subjective wellness, injury, and training load. Additionally, the dashboard enables time series forecasting and correlation analysis, making it a valuable resource for athletes, coaches, team staff and researchers.

### 2.5.1 Analyses and Visualisations

The dashboard currently includes seven types of analyses and visualisations divided between two pages: one for individual players and one for the entire team [10]. Three types of analyses and visualisations are provided on the player information page using an anonymised UUID for identification. The components from the player page are illustrated in Figure 2.5. The first type of analysis is readiness forecasting, which uses an Autoregressive Moving Average (ARIMA) model with exogenous regression to predict a player's readiness to train for the next days based on data from past days.

The second analysis is an overview of a player's wellness parameters,


Figure 2.6: Soccer Dashboard team analyses and visualisations [10]

visualised in a spider or radar chart. This chart aggregates wellness values over time and allows coaches to select between different time ranges, such as last week, last fortnight, last month or last year. This allows coaches to observe recent or overall wellness and identify increasing or decreasing values over time and how different parameters may affect each other.

The third analysis overviews a player's training load over a specific time range, specifically their ACWR values. This is displayed in a bar chart using blue, grey, and red to identify low, normal, or high training load days. The identification of high training load days is based on the upper 75% quartile of the athletes' past data, while the identification of low training load days is based on the lower 25% quartile. ACWR values combine the player's perceived exertion and exhaustion in the current week with the values from the last 42 days, indicating fatigue and fitness. This allows coaches to see the players' progress.

The team information page includes four modules for analysing and visualising data related to the entire team, as shown in Figure 2.6. The first module is a quartile-based representation of wellness and training load metrics collected over a chosen time period across all players for a given team. The module provides trainers with an overview of the team's overall state, which can be used to reflect upon training regimes based on the ranges of values for specific parameters.

The following module is a summary of all injuries that have occurred within a team. The visualisation separates injuries by location and severity, allowing coaches to identify the most common injury locations and their

severity.

The third module is a training load overview for the entire team, which is presented as a table with the mean values and standard deviation for each training load metric per player. Both tables can be downloaded as CSV files. This functionality is provided to allow users to do their own analysis of a team.

Lastly, a module displays a correlation matrix of the wellness parameters and the training load metrics. The matrix illustrates the linear relationships between the various metrics, including those among the wellness parameters and between the training load parameters. This correlation matrix can be used to understand how the metrics may influence one another and potentially predict future values.

### 2.5.2 Visualising Injury Analysis

Visualising injury analysis aids in a better understanding of the injuries, their causes and how to prevent them. By presenting injury data in a dashboard format, coaches and players can identify patterns and adjust their training programs to reduce the risk. Similarly, medical staff, including physiotherapists and doctors, can use the dashboard to organise treatment plans and assist in recovering injured players.

In addition to its usefulness for those directly involved with soccer teams, injury analysis dashboards can also be valuable for researchers. By providing access to detailed data and analyses, researchers can gain insights that can inform their work and contribute to a broader understanding of injury prevention and management.

The dashboard also allows for interactivity so that users can choose and experiment with date ranges, variables and similar. This is useful because they can specify the range of the analysis and see how it varies for different players or parameters.

## 2.6 Injury Risk

Sports injuries are a significant problem for athletes at all levels. Injuries can lead to decreased performance, missed games or seasons, and in severe cases, can even end an athlete's career [8]. Therefore, injury prevention is a critical aspect of sports training and competition.

Injuries in sports can be acute, such as a muscle strain or broken bone, or chronic, such as overuse injuries like tendonitis [1, 58]. Athletes in contact sports such as football, hockey, basketball and soccer are at higher risk of acute injuries due to the nature of the sport, but all athletes are at risk of both acute and chronic injuries [54].

The impact of sports injuries is not limited only to athletes. Injuries can significantly impact teams and organisations, with costs associated with medical treatment, rehabilitation, and replacement of injured players [62]. Additionally, injuries can affect team morale and cohesion, leading to decreased performance and decreased success on the field.

To prevent injuries, it is essential to have a comprehensive injury prevention program that includes proper warm-up and stretching techniques, adequate rest and recovery time, proper equipment and technique, and monitoring of training load [20]. Injury prevention programs can significantly reduce the risk of injuries and help athletes perform at their best.

Identifying and understanding the risk factors associated with sports injuries is also essential. Risk factors include age, previous injuries, training load, and technique, among others [33]. By identifying these risk factors, coaches and trainers can develop strategies to reduce the risk of injury and keep athletes healthy and performing at their best.

A factor commonly used to assess injury risk is the training load metric ACWR[63]. ACWR is the ratio between the athlete's current and chronic training load from the last month or so, reflecting the athlete's fatigue and fitness levels [4]. Higher ACWR values reflect a higher acute training load and lower chronic load, meaning a sudden increase in the training load. An ACWR value of 1 would mean no changes in the training load. Lower ACWR values reflect higher chronic and lower acute loads, meaning that the athlete's training load is lower than usual.

Tim J. Gabbett [25] proposed an injury risk model based on ACWR values and recommended maintaining ACWR values between 0.8 and 1.3, referred to as the "sweet spot" for minimised injury risk, as illustrated in Figure 2.7. This injury risk model is the consensus model for the International Olympic Committee (IOC) and has been used to instruct training recommendations [57].

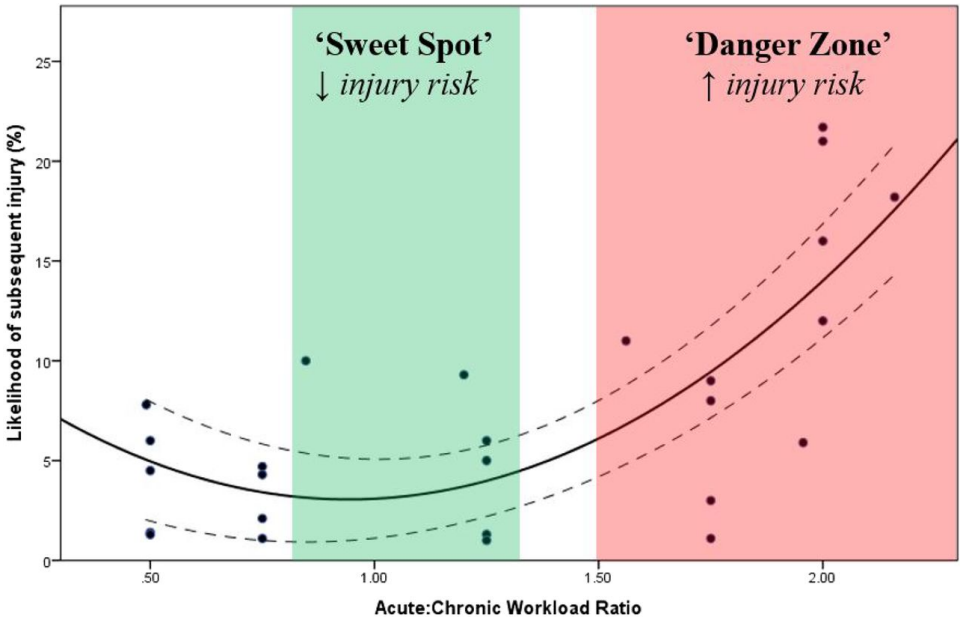



Figure 2.7: ACWR "sweet spot" [25]

However, recent discussions by Wang et al. [63] have highlighted limitations to using ACWR as a predictor of injury risk and advocated for newer strategies. One of their points is that the ratio between acute

load and chronic load may not accurately reflect changes in load, as acute load could be high for various reasons, such as competition or increased training volume, while chronic load could remain unchanged. Second, the weighting of each week in the chronic load is the same, so the load from four weeks prior would have the same impact on the ACWR as the load from the last week. This may not accurately reflect an athlete's current state.

Additionally, the model oversimplifies the relationship between training load and injury risk. There may be a "sweet spot" where injury risk is lowest, but this is not necessarily the same for all sports and athletes. Injury risk may vary based on the athlete's characteristics and the type of sport they participate in.

According to Bahr and Holme [5], looking into multiple injury risk factors is essential because no single factor can fully explain the complex nature of sports injuries. While ACWR can be helpful, it only provides a snapshot of an athlete's training load. It does not account for other factors, such as biomechanics, nutrition, sleep, and psychological factors, which can all contribute to injury risk. Examining multiple risk factors provides a more comprehensive understanding of an athlete's injury risk, which can lead to better injury prevention strategies.

## 2.7 Overview of Related Work

Researchers have explored various strategies in sports injury prevention, including using ACWR, machine learning and survival analysis. In this section, we will present some notable implementations and experiments undertaken in this field that are relevant to our research in this thesis.

**Predicting Injury Risk Over Changes in Physical Activity in Children Using the Acute:Chronic Workload Ratio** Wang et al. [64] conducted a study investigating the prognostic relationship between changes in activity, measured by ACWR, and injury in children. They used training load and injury data collected from 1660 children aged 6–17 years over the course of 3.8 years. The physical activity of the children refers to physical education classes. To model the relationship between the uncoupled 5-week ACWR and injury, defined as patient-reported musculoskeletal pain, they used a Generalised Additive Mixed Model (GAMM).

They found minimal differences in risk for ACWR values between 0.4 and 1.0. There were gradually lower injury risks with lower ACWR values approaching 0. Minimal differences in risk were predicted with ACWR values ranging from 1.0 to 1.3. For ACWR higher than 1.3, there were gradually higher risks.

This is relevant to our research as we also use ACWR data in our approach. As they found that higher ACWR values above 1.3 increase the risk of injury, this supports the suggestions by Gabbett [25] that ACWR values above 1.3 increase risk. However, their results also showed that lower ACWR values decrease injury risk, meaning the injury risk model by

Gabbett does not apply. The injury risk model suggests a higher risk for lower ACWR values below 0.8.

As their study may prove helpful in our thesis, there are some differences that could result in other findings in our research. For instance, their research was conducted using data from children and physical education classes, but the data we use is collected from adult soccer players. Children may react differently to changes in ACWR values than adult athletes, and adults may have different training and recovery needs than children. Additionally, training load and injury risk in physical education classes might completely differ from that in soccer. Our research also investigates the relationship between wellness factors and injury risk while they solely focused on ACWR. Another difference is that they used GAMM for analysis, which is a different statistical approach than survival analysis. These two types of statistical analysis might provide different results.

**Athlete-Customised Injury Prediction** The Bio-Engineering Department at the University of Louisville developed a framework for predicting non-contact injuries resulting from intense workouts by utilising objective data from wearable devices and subjective data from questionnaires regarding training load [45]. The framework was tested on 21 soccer players and aimed to estimate the load on each athlete, which can be both internal, such as heart rate, and external, such as duration of workout and the number of jumps. Their framework predicts injuries and provides a guide for athlete-specific factors that contribute to increased load, thus helping in injury prevention for individual athletes or groups.

To achieve this, the researchers first built datasets for machine learning techniques by selecting the most relevant features, which were 65 wearable GPS-equipment measurement features. They then grouped these features into non-injured and injured samples, with injured samples containing records that occurred within one week before the injury. They performed machine learning analysis using KNN and K-means classification on each feature. This resulted in a framework that provides a probability mapping of injury factors specific to each player.

This is relevant to our research as they identify injury risk factors and use data from adult soccer players. However, they only focus on objective and subjective training load, and we focus on both training load and wellness in our research. Additionally, they only use records from within a week before the injury, and we use all data prior to injuries to identify more long-term factors.

While their approach using machine learning techniques is useful for predicting injuries based on objective and subjective data, survival analysis may provide a more appropriate method for identifying injury risk factors. Survival analysis accounts for censored data, models the time-to-event outcome, and estimates hazard ratios to identify factors associated with an increased risk of injury [48]. Using survival analysis, we can gain more insights into the underlying mechanisms contributing to injury risk and develop more targeted injury prevention strategies.

**Subjective Well-Being and Training Load Predict In-Season Injury and Illness Risk in Female Youth Soccer Players** Researchers at the Department of Orthopedics at the University of Wisconsin School of Medicine and Public Health conducted a study evaluating the effects of training load and well-being on injury and illness risk in female adolescent soccer players [66]. Using Poisson regression modelling the number of injuries that occurred in a given time period as a function of various predictors, they could predict daily injuries and illnesses. They found that lower mood and higher acute training load were associated with increased injury risk on the day of the injury. Higher acute training load suggest higher ACWR values, meaning that their results match both Wang et al. and Gabbett's recommendation on higher ACWR increasing injury risk.

Similar to our research, they use training load and wellness factors to assess injury risk. They also collected data on female soccer players, but unlike our data, they focused on adolescent soccer players. As mentioned earlier in this section, adult athletes may react differently to these factors than younger athletes. They also focus more on the acute factors, such as mood on the day of the injury, but we focus more on the long-term effects leading up to the injury. Additionally, they use a different method of analysis, which could yield different results than survival analysis.

**Time-To-Event Analysis for Sports Injury Research Part 1: Time-Varying Exposures** Nielsen et al. [48] conducted a study discussing statistical methods appropriate for the complex analysis of time-varying variables, such as changes in training load and injury-related outcomes. The advanced statistical methods they discuss are time-to-event analysis, or in other words, survival analysis. They state that the investigator must analyse exposure variables that change over time, such as changes in training load. They also note that very few studies have included time-varying exposures, such as training load, and time-varying effect-measure modifiers, such as previous injury, biomechanics, sleep and stress, when studying the cause of sports injuries.

As they point out, based on a systematic review of articles examining the relationship between training load and sports injury, the most common analytical approaches were $X^2$ tests and logistic regression. Because it is not possible to include time-varying exposures in traditional logistic regression models or $X^2$ tests unless the time to injury is discretised, they suggest that one should investigate survival analysis to assess the relationship between changes in training load and injury. There have been few documented approaches using survival analysis for sports injuries. Specifically, based on the review, less than 10% of all results in the identified studies were based on survival analysis.

This study is relevant to our research as it highlights the importance of using survival analysis to study the relationship between time-varying exposures and injury outcomes. As very few studies have included time-varying variables when studying the cause of sports injuries, we can address this gap by identifying injury risk factors using time-varying

variables.

**Monitoring of Sport Participation and Injury Risk in Young Athletes** Laurent Malisoux et al. [39] conducted a study examining the relationship between sports participation patterns and injury risk in young athletes participating in team, individual and racket sports using the CPH model. This is one of the studies mentioned by Nielsen et al. as the few survival analysis approaches for estimating injury risk. They collected subjective training load and injury data from 154 young elite athletes over the course of 41 weeks using the electronic system TIPPS (Training and Injury Prevention Platform for Sports) for reporting. The athletes reported RPE (Rating of Perceived Exertion) for each sport session, which was used to compute weekly load, monotony and strain.

They found that team sports had the highest injury risk and that short-term modifications in intensity were associated with injury risk, such as higher intensity in the last week compared to the last 4 weeks. In other words, a higher acute training load and lower chronic load, resulting in higher ACWR, are associated with injury risk. This supports the concept of higher ACWR and injury risk being related, as discussed in Section 2.6. This also supports the use of survival analysis to analyse the relationship between training load and the injury event, as Nielsen et al. mentioned in the previous paragraph.

This is relevant to our research as it provides evidence for using survival analysis in investigating the relationship between training load and injury risk. Additionally, their finding on higher ACWR increasing injury risk is relevant to us as we also use ACWR in our experiments.

On the other hand, they only focus on training load in young athletes, whereas we include wellness metrics and use data from adult athletes. Therefore, we extend the scope of Malisoux et al.'s research by introducing wellness factors and data from adult athletes.

As presented in this section, various studies related to sports injury assessment use approaches such as ACWR, machine learning, and survival analysis. The studies use training load or wellness factors for injury risk assessment but differ from our research in focusing on long-term or acute factors, age groups, and methodology. We can use their findings to support our research and expand by including more variables, using survival analysis for our methods and data from adult female soccer players.

## 2.8 Chapter Summary

In this chapter, we provide a comprehensive overview of athlete monitoring and the collection and utilisation of athlete monitoring data by the FFRC from elite female soccer players. Specifically, we present the pmSys system and the SoccerMon dataset, the primary data sources for our experiments in this thesis. We also present Soccer Dashboard, a visualisation and analysis tool using the SoccerMon dataset.

We discuss ACWR as a commonly used measurement for injury risk and highlight the need for newer strategies and the consideration of multiple risk factors to improve injury prevention.

Moreover, we examine related work in injury prevention, including using ACWR, machine learning, and survival analysis techniques. By analysing and evaluating these related studies, we identify gaps in the current understanding of injury risk assessment and discuss how our research aims to address these shortcomings. This chapter aims to provide a deeper understanding of the motivation behind our methods, which we describe in detail in the subsequent chapter.

# Chapter 3

# Methodology

The previous chapter provided contextual information and presented previous research relevant to this thesis. The groundwork laid out in that chapter will prove essential to our research.

In this chapter, we introduce the concept of survival analysis and the techniques we use in this thesis. We present the survival models employed in our research and how these are used for validating data distribution and estimating the significance of covariates. We also present the data structures needed for these models. Next, we describe how regularisation can be used for feature selection and address the potential issues of overfitting and multicollinearity. We present our methodology for evaluating the goodness of these models using estimates such as the Bayesian Information Criterion (BIC) and the Concordance Index (C-index) from cross-validation. Finally, we outline the different tools we use to implement our methods.

## 3.1 Survival Analysis

Survival analysis is a statistical technique used to analyse time-to-event data. The "event" can be any occurrence of interest, such as a death, failure, or in our case, injury [13]. In contrast to other statistical techniques that only consider the time of an event, survival analysis also considers that some events may not have occurred by the end of the study, referred to as *censoring*. Survival analysis can also be used to identify risk factors associated with the event. For the study, one typically observes a group of individuals and records occurrences of events or "deaths" over time. The time up until the specific event is referred to as the *duration* interval, starting from the first observation, their "birth".

In this section, we provide an overview of some of the fundamental survival analysis concepts essential for understanding our methods and experiments. We discuss the different types of events, the importance of censoring and how it relates to survival analysis and introduce the concepts of survival, hazard, and cumulative hazard functions.

### 3.1.1 Events

Events in survival analysis refer to any occurrence of interest, such as death, failure, disease or injury [36]. In this thesis, we use two types of events: first events and recurrent events.

First events are the initial occurrence of an event in a subject during the observation period, such as the first occurrence of a heart attack or stroke. The focus of the analysis is typically on the time prior to the first event for each individual. The analysis assumes that once individuals experience the first event, they are no longer at risk for that particular event.

Recurrent events refer to the occurrence of the same event multiple times in the same individual [3]. For example, in a study of migraine headaches, a recurrent event could be the number of headache episodes experienced by each participant. Analysis using recurrent events models the time between successive events, rather than just the time to the first event, and accounts for the fact that an individual can experience multiple events.

### 3.1.2 Censoring

Individuals may not experience the event of interest during the observation period for various reasons, such as dropping out of the study or the observation period ending before the event occurs [13]. These individuals are considered censored. Right-censoring is the most common type of censoring in survival analysis. The duration interval ranges from the start to the end of the observation period, also known as the time of censoring. Figure 3.1 illustrates an example of right-censoring in a study over 20 days. Day 0 represents the start of the study, and the stippled line at day 20 represents the end of the study. The red subjects are individuals who experienced the event, the red dot, during the study. The blue subjects are censored individuals who did not experience the event during the study. They may experience the event after the study, as seen in individual 2, or never experience it, as seen in individual 6. The censored individuals are on the right side of the end of the study, marking the explanation of why they are called right-censored.

Left-censoring and interval-censoring are the two other types of censoring that may occur in survival analysis. Left-censoring occurs when the true event time is unknown, but it is known that it occurred before a specific time [36]. This type of censoring is often encountered in studies where the observation period begins after the event of interest. As a result, some individuals have already experienced the event before the study starts. The duration interval for left-censored observations is from the unknown event time to the end of the observation period.

Interval censoring is when the true event time is only known to have occurred within a certain time interval [36]. This type of censoring occurs when the event is only observed at certain time intervals or when the exact event time is uncertain. For example, in a cancer screening study, the exact time when the disease started may not be known, but it can be determined

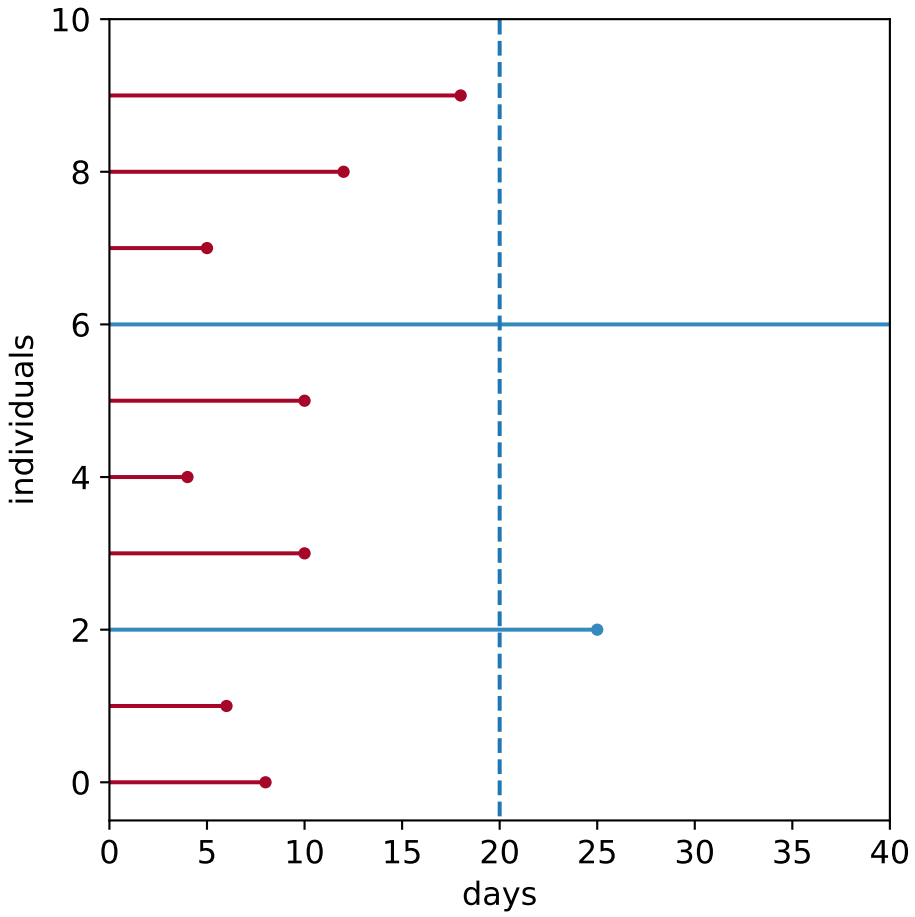



Figure 3.1: Censoring

that it occurred within a certain time period. The duration interval for interval-censored observations is from the lower bound of the interval to the upper bound.

Censoring is essential in survival analysis because it reflects the reality that not all individuals in a study will experience the event of interest during the observation period [18]. By accounting for censored observations, it is possible to estimate the survival and hazard functions more accurately and make more reliable predictions about the risk of experiencing the event. Censoring also allows for more efficient use of resources in conducting survival studies, as it is often not feasible or practical to follow individuals until the event of interest occurs for everyone.

### 3.1.3 Survival Function

In survival analysis, the survival function is a fundamental concept used to model the probability of an individual surviving up to a given time without experiencing the event of interest [36]. The survival function is denoted by

$S(t)$, where $t$ represents the time variable of interest.

The survival function can be defined as the probability that an individual will survive beyond time $t$, given that they have survived up to time $t$. Mathematically, the survival function can be expressed as:

$$S(t) = P(T > t) \tag{3.1}$$

where $T$ is a random variable representing the time to the event of interest. The survival function can also be interpreted as the proportion of individuals in a population who have not experienced the event of interest up to time $t$.

The survival function is different for every survival model because it depends on the model's underlying assumptions. For example, the CPH model assumes that the hazard function is constant over time, which leads to a specific form of the survival function [36]. In contrast, the parametric survival models assume a specific functional form for the hazard function, which leads to a different form of the survival function. Therefore, the choice of survival model affects the form of the survival function and the interpretation of the results.

The survival function is essential because it allows us to estimate the proportion of individuals that survive past a given time and compare survival probabilities between different groups or treatments. It is also used to estimate other important quantities in survival analysis, such as the hazard function and the cumulative hazard function.

### 3.1.4 Hazard and Cumulative Hazard Function

The hazard function and cumulative hazard function are both important survival analysis concepts used to model the probability of an event occurring over time [36].

The hazard function, denoted by $h(t)$, is the immediate rate at which events occur at time $t$, given that the individual has survived up to time $t$. In other words, the hazard function represents the probability that an event will occur in a small interval of time, given that the individual has not experienced the event up to that point in time. The hazard function is a key component in survival analysis, as it provides a way to model the risk of an event occurring over time.

Mathematically, the hazard function can be expressed as:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T \leq t + \delta t | T > t)}{\delta t} \tag{3.2}$$

where $T$ represents the time-to-event variable of interest, and the limit is taken as the time interval $\delta t$ approaches zero [36]. This expression represents the conditional probability that an event occurs in a small interval of time $\delta t$, given that the individual has survived up to time $t$. Dividing by $\delta t$ and taking the limit as $\delta t$ approaches zero results in the instantaneous risk or probability of the event occurring at time $t$.

The cumulative hazard function, denoted by $H(t)$, is the cumulative sum of the hazard function up to time $t$. In other words, the cumulative

hazard function represents the total amount of risk that an individual has experienced up to a particular point in time. The cumulative hazard function is also an important concept in survival analysis, as it provides a way to model the overall risk of an event occurring over time.

Mathematically, the cumulative hazard function can be expressed as:

$$H(t) = \int_0^t h(z)dz \tag{3.3}$$

where $h(z)$ is the hazard function at time $z$, and the integral is taken from 0 to t [36].

The hazard function and cumulative hazard function are used to model the survival distribution in survival analysis. The survival function is the complement of the cumulative hazard function because it represents the probability that an individual will survive up to time t without experiencing the event of interest.

## 3.2 Survival Analysis Models

As discussed in Sections 3.1.3 and 3.1.4, the survival function estimates the probability of an individual surviving beyond a given time point. The hazard function estimates the instantaneous rate at which events occur at a given time, given that the individual has survived up to that point. Survival analysis offers several models for estimating survival and hazard. In this section, we present the models we use in our research.

Survival analysis models can be broadly classified into univariate and multivariate [13]. Univariate models assess the relationship between a single variable and the outcome of interest [28]. In survival analysis, this is the relationship between duration intervals and the survival outcome [36]. These models estimate the survival or hazard function solely based on durations and events recorded during an individual's observation period. Our research uses univariate models to validate if our data can be applied to survival analysis. We do so by examining the shape of the survival curve and comparing it to our distribution of durations and events. The curve can indicate if there are any outliers and if the distribution of event times is symmetric or skewed. Therefore, if the data produces a good survival curve, it suggests that the data can be effectively analysed using survival analysis. The univariate models employed in our thesis include Kaplan-Meier, Weibull, and Piecewise Exponential.

Multivariate models consider the relationship between the survival time of individuals and multiple covariates, which are typically factors like age, gender and medication [13]. These models incorporate duration intervals, events, and additional covariates to determine how these covariates influence the survival or hazard function. We use multivariate models in our research to estimate the effects of the covariates on injury outcomes. We use the Cox Proportional Hazards Model (CPH) and its extension, the Cox Time-Varying Model (CTV), as our multivariate models.

### 3.2.1 Kaplan-Meier

The Kaplan-Meier model is a univariate model for estimating the survival function [36]. It calculates the survival probability for each observed survival time and then multiplies them together to obtain an overall survival function estimate.

Specifically, the Kaplan-Meier estimator of the survival function $S(t)$ at time $t$ is presented in Equation 3.4.

$$S(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i} \tag{3.4}$$

where $t_i$ is the i-th ordered survival time, $d_i$ is the number of events that occur at that time, and $n_i$ is the number of individuals still at risk just prior to time $t_i$ [36].

The Kaplan-Meier model is a common method for analysing survival data because it is simple and does not require any assumptions about the underlying distribution of survival times, making it applicable to a wide range of datasets [29]. Also, the Kaplan-Meier curve is easy to interpret, and we can use this to compare it to our data distribution and validate our data for survival analysis.

### 3.2.2 Weibull

The Weibull model is a univariate model that estimates the hazard function as a function of time using a Weibull distribution. The Weibull distribution is a flexible distribution that can be used to model various shapes for the hazard function, including increasing, decreasing, and constant hazards [36].

The Weibull model assumes a shape parameter ($\beta$) that determines the shape of the hazard function and a scale parameter ($\lambda$) that determines the overall hazard level. The hazard function $h(t)$ for the Weibull model is given by:

$$h(t) = \frac{\beta}{\lambda} \left( \frac{t}{\lambda} \right)^{\beta-1} \tag{3.5}$$

where $t$ is the time to the event of interest. The shape parameter $\beta$ determines the shape of the hazard function, with values greater than 1 indicating an increasing hazard rate over time, values less than 1 indicating a decreasing hazard rate over time, and a value of 1 indicating a constant hazard rate over time. The scale parameter $\lambda$ determines the overall hazard level, with larger values indicating a lower hazard rate.

We can use the Weibull model similarly to Kaplan-Meier by comparing its curve to our data distribution. However, the model's flexibility in modelling the hazard shape is better suited for situations where the hazard rate is not constant, which could result in more precise estimates compared to Kaplan-Meier [40].

### 3.2.3 Piecewise Exponential

The Piecewise Exponential model is a univariate model used to estimate the hazard function over time, assuming that the hazard function changes at certain points, or knots, along the time axis [49]. Each interval between two knots is characterised by a constant hazard rate, allowing the model to account for changes in hazard over time.

Mathematically, the hazard function for the Piecewise Exponential model can be expressed as:

$$h(t) = \begin{cases} \lambda_0, & \text{if } t \leq \tau_0 \\ \lambda_1, & \text{if } \tau_0 < t \leq \tau_1 \\ \lambda_2, & \text{if } \tau_1 < t \leq \tau_2 \\ \ldots \end{cases} \tag{3.6}$$

where $h(t)$ is the hazard function at time $t$, $\lambda_i$ is the constant hazard rate for interval $i$, and $\tau_i$ represents the knots along the time axis, separating the intervals [49].

The Piecewise Exponential model is beneficial for modelling non-proportional hazards, where the hazard rate changes significantly over time. This can occur, for example, in medical studies where a treatment effect diminishes over time or in studies of environmental exposure where the risk of an event decreases as exposure decreases.

One drawback of the piecewise exponential model is that it assumes that the hazard rate is constant within each interval, which may not be the case in reality.

The Piecewise Exponential model can be used similarly to the Kaplan-Meier and Weibull models by comparing its survival function to the injury distribution. However, it may be a better option in situations with significant changes in the hazard rate over time.

### 3.2.4 Cox Proportional Hazards Model

The CPH model is a multivariate model that investigates the relationship between the survival time of individuals and one or more covariates [14]. The model is considered semi-parametric because it uses a baseline hazard function that is not fully specified but estimated non-parametrically. At the same time, the effect of covariates is assumed to be proportional and modelled parametrically [69].

Mathematically the CPH model is written as:

$$h(t|x) = h_0(t) \exp\left(\sum_{i=1}^{n} b_i x_i\right) \tag{3.7}$$

where $h(t|x)$ is the hazard function at time $t$ for a given set of $n$ covariates $x$ [13]. The impact of the covariates is measured by the size of their corresponding coefficients $b$. $h_0(t)$ is the baseline hazard function at time $t$ and is the value of the hazard if all the covariates are equal to zero.

The CPH model assumes that the relative hazard between two individuals remains constant over time, known as the *proportional hazards assumption* [36]. The assumption implies that the difference in the hazard rate between two individuals remains the same at any time. The assumption is necessary because violating it can lead to biased and unreliable results [37]. If the effect of the covariate on the hazard change over time, the estimated hazard ratio may not accurately reflect the true relationship between the covariate and the hazard.

The effect of the covariates can be investigated by analysing the significance of their coefficients and hazard ratios [13]. The hazard ratio measures the change in the hazard rate for a one-unit change in a covariate, in other words, the effect of the covariate on the event outcome. The hazard ratio is calculated as the exponential of the coefficient, $\exp(coef)$. The coefficient is derived by the logarithm of the hazard ratio, $log(HR)$. The larger the hazard ratio, the more effect the covariate has on the outcome. A hazard ratio of 1 indicates no difference in hazard rate [36]. In contrast, a hazard ratio greater than 1 indicates a higher hazard rate, and a hazard ratio less than 1 indicates a lower hazard rate. For instance, if the covariate age has a coefficient value of 0.03, this would mean that the hazard rate increases by 3% with each unit increase in age. The hazard ratio is $\exp(0.03) = 1.03$, which is close to 1, meaning age has a small effect on the hazard rate. For a higher coefficient of 2.0, on the other hand, the hazard rate increases by 200% for each unit increase, meaning it doubles, resulting in a hazard ratio of $\exp(2.0) = 7.39$. Negative coefficients indicate that increasing the covariate value decreases hazard and increases survival. For instance, if a covariate has a coefficient of -0.9, this would mean that the hazard rate decreases by 90% for each unit increase in that covariate. The hazard ratio would be $\exp(-0.9) = 0.41$. Therefore, significantly negative coefficients have an essential impact on outcomes in terms of survival.

A *confidence interval* is a statistical range of values in which an estimate is likely to be located based on a given set of data and a chosen level of confidence [7]. The interval consists of the estimate's mean and lower and upper bounds, meaning the variations of the estimate in the data. The level of confidence refers to the probability that the true estimate falls within the calculated interval. For example, a 95% confidence interval means that 95 out of 100 times, the estimate falls between the upper and lower bounds. We can use confidence intervals with the coefficients described above as estimates to evaluate their reliability. A large confidence interval suggests that the sample size is relatively small, the variation in the data is large, or both [52]. It also suggests that the estimated coefficient may not be very reliable or precise and that more data may be needed to estimate the true value of the coefficient better.

The CPH model is a valuable tool in survival analysis for estimating the effect of factors on the event outcome and is helpful in our research for identifying significant injury risk factors. By analysing the coefficients and their confidence intervals, we can also evaluate the certainty of the estimated effects.

### 3.2.5 Cox Time-Varying Model

The CTV model is an extension of the CPH model, as it assesses the impact of time-varying covariates on the hazard function [36]. The model includes covariates that change over time, such as a change in treatment, stress levels or training load. These time-varying covariates are included in the model by defining them as time-dependent variables, which are updated at each observation time. Time-independent variables, on the other hand, do not change over time or during the observation period and are variables such as age and gender.

The general form of the CTV model can be written as:

$$h(t|x) = h_0(t)exp\left(\sum_{i=1}^{n} b_i x_i(t)\right) \tag{3.8}$$

where $h(t|x)$ represents the hazard function at time $t$ given the values of the time-dependent covariates $x(t)$. $h_0(t)$ is the baseline hazard function at time $t$, and $b$ are the coefficients for the time-dependent covariates.

The coefficients, hazard ratios and confidence intervals from the CTV model can be interpreted as described in the previous section about the CPH model.

The CTV model is useful when covariates change over time, which might affect the hazard rate of the event of interest [70]. For example, in medical studies, a patient's treatment may change during the study, and the effect of the treatment on survival may change over time. In this case, the CTV model can provide more accurate and informative results than the standard CPH model.

However, the CTV model violates the proportional hazards assumption because it allows the hazard ratio to vary over time, unlike the traditional CPH model [36]. As time-dependent covariates are allowed to have different effects on the hazard function at different time intervals, the assumption of a constant hazard ratio is violated. Therefore, the CTV model might lead to biased and unreliable results [37].

The time-varying model can be helpful for our research as it can assess the impact of time-dependent covariates on the hazard function. As we use time-series data and have time-dependent covariates, the model can provide more accurate and informative results, as it can reflect changes in the real world. However, we must be aware of the potential limitations of the model, as it violates the proportional hazards assumption.

## 3.3 Data Structure

In order to utilise the models presented in Section 3.2, the data must be structured in a specific format. This section will describe how we build data structures with censoring and calculated duration intervals for univariate and multivariate survival analysis models.

As described in Section 3.1.2, data entries missing events of interest are still included in the analysis, as they are censored. Censored data is

represented in a specific format. The format is referred to as *counting process format* or *time-to-event format* [36]. The variable $C_i$ is binary and denotes whether the observation is censored. Hence, $C_i$ is true if an event occurs and false if not.

Durations are calculated as the number of days between two observation points. The duration start point is usually the first observation. The duration endpoint is either an observed event or the last observation point, meaning it is right-censored. The variable $T_i$ denotes the number of days in the duration interval.

### 3.3.1 Structure for Univariate Models

Table 3.1 shows how data can be structured for univariate models such as Kaplan-Meier, Weibull and Piecewise Exponential. The individuals are soccer players, and the event of interest is injury. A duration lasts from the first day of observation to either the first injury occurrence or the last day of observation if no injury occurred. The duration column represents $T_i$, and the event column represents $C_i$. For players 1 and 4, their event is set to true, meaning they experienced injuries. For players 2 and 3, their event is false, meaning they did not have an injury in the observation period, and are therefore censored. The observation period for player 2 lasted 119 days, but for player 3, it lasted 543 days. The observation period does not look the same for all individuals if they miss days with observation. For this example, it could indicate that player 2 missed days of reporting or dropped out of observation, seeing that player 2 has fewer observation days than player 3.

| **player_name** | **duration** | **event** |
|---|---|---|
| player 1 | 37 | True |
| player 2 | 119 | False |
| player 3 | 543 | False |
| player 4 | 52 | True |

Table 3.1: Example data structure for univariate models

### 3.3.2 Structure for Multivariate Models

As described in Section 3.2, multivariate models, such as the CPH model, use factors from the data to evaluate their impact on survival probability. Using the previous example, Table 3.2 displays how data can be structured with covariates such as ACWR and readiness. The values presented in the table correspond to either the day of the event or the day of the last observation. For example, on the day of their injury, player 1 had an ACWR of 0.8 and a readiness score of 8. On the day of their last observation, player 2 had an ACWR of 0.5 and a readiness score of 8.

When dealing with covariates that vary over time, it is important to consider all the values within the duration interval, not just on the day of

| player_name | duration | event | acwr | readiness |
|---|---|---|---|---|
| player 1 | 37 | True | 0.8 | 8 |
| player 2 | 119 | False | 0.5 | 8 |
| player 3 | 543 | False | 1.5 | 2 |
| player 4 | 52 | True | 4.1 | 5 |

Table 3.2: Example data structure for multivariate models

the event. This is because changes in the covariate prior to the event might impact the outcome. One approach to address this is the CTV model, as described in Section 3.2.5. This model requires a different data structure than the previous examples. Instead of having one entry per duration, this format has an entry for each state change made during the duration interval, with corresponding covariate values for that state observation [36]. The structure includes a start variable indicating the day of the last observation, a stop variable indicating the day of the current observation, and covariate values from the current observation day.

Table 3.3 illustrates an example of this time-varying data structure. It includes daily observations of players 1 and 2 from the previous examples, with the event set to false until the day an injury occurs, as seen for player 1 on day 36. For player 2, who is censored, the event remains false.

| player_name | start | stop | event | acwr | readiness |
|---|---|---|---|---|---|
| player 1 | 0 | 1 | False | 2.2 | 4 |
| player 1 | ... | ... | ... | ... | ... |
| player 1 | 36 | 37 | True | 0.8 | 8 |
| player 2 | 0 | 1 | False | 1.5 | 6 |
| player 2 | ... | ... | ... | ... | ... |
| player 2 | 118 | 119 | False | 0.5 | 8 |

Table 3.3: Example data structure for multivariate time-varying model

On the other hand, while the CTV model allows for the modelling of time-varying covariates, it does not meet the proportional hazards assumption, as discussed in Section 3.2.5. Moreover, the data structure required for this model is more complex and involves generating a larger table.

Alternatively, a more straightforward approach to incorporating time-varying covariates is to calculate the mean value of the covariates over the entire duration period. This provides an overall measure of the covariate values during that period and is compatible with the CPH model, which satisfies the proportional hazards assumption.

Table 3.4 displays an extension of Table 3.2, including all the ACWR and readiness values from each player's duration interval. Table 3.5 shows the averaged covariate values from Table 3.4.

Recurrent events, as explained in Section 3.1.1, provide another option to incorporate the time-varying nature of the covariates. Including all events for an individual and then calculating the mean covariate values for

| player_name | duration | event | acwr | readiness |
|---|---|---|---|---|
| player 1 | 37 | True | [2.2 ... 0.8] | [4 ... 8] |
| player 2 | 119 | False | [1.5 ... 0.5] | [6 ... 8] |
| player 3 | 543 | False | [0.8 ... 1.5] | [5 ... 2] |
| player 4 | 52 | True | [1.1 ... 4.1] | [7 ... 5] |

Table 3.4: Example data structure for multivariate models using all values

| player_name | duration | event | acwr | readiness |
|---|---|---|---|---|
| player 1 | 37 | True | 1.5 | 6.0 |
| player 2 | 119 | False | 1.0 | 7.0 |
| player 3 | 543 | False | 1.2 | 3.5 |
| player 4 | 52 | True | 2.6 | 6.0 |

Table 3.5: Example data structure for multivariate models using mean values

all the durations captures more covariate values over time and provides more data for analysis. However, when multiple events exist for the same individual, the CPH model's assumption of independent events may not hold. To address this, a binary covariate, such as prior injury, can be introduced to model the correlation between events. Introducing this variable also allows us to assess how much prior events affect the outcome and compare it with the other covariates.

Table 3.6 illustrates an extension of Table 3.5, where recurrent injuries and the prior injury variable are included. Player 1 has two injuries, one on day 37 and another 45 days later. The prior injury variable is false for the first injury and true for the second injury. The averaged covariate values also change from duration to duration.

| player_name | duration | event | prior_injury | acwr | readiness |
|---|---|---|---|---|---|
| player 1 | 37 | True | False | 1.5 | 6.0 |
| player 1 | 45 | True | True | 1.0 | 6.5 |
| player 2 | 119 | False | False | 1.0 | 7.0 |
| player 3 | 543 | False | False | 1.2 | 3.5 |
| player 4 | 52 | True | False | 2.6 | 6.0 |
| player 4 | 38 | True | True | 3.7 | 4.5 |

Table 3.6: Example data structure for multivariate models using recurrent events

## 3.4 Feature Selection

Feature selection is selecting a subset of relevant variables from a larger set of available features in a dataset [50]. The goal is to identify the most important and informative variables that can effectively predict the target

variable while discarding the irrelevant, redundant or noisy variables.

In our research, we use *regularisation* with the CPH model for feature selection, as our goal is to identify the factors with the most impact on the outcome, which is injury. Regularisation is a widely-used technique for feature selection, which involves adding a penalty term to the cost function of the model to eliminate irrelevant or noisy covariates [50]. Apart from feature selection, regularisation also has the advantage of preventing overfitting and mitigating multicollinearity issues.

Overfitting occurs when a model is too complex and has too many variables relative to the training data available. This can lead to poor performance on new, unseen data [68]. Regularisation addresses this issue by eliminating irrelevant covariates and reducing the number of covariates in the model.

Multicollinearity occurs when multiple input variables are highly correlated, leading to unstable estimates of the regression coefficients [2]. Regularisation techniques can be used to shrink the coefficient estimates or exclude them altogether to mitigate the impact of multicollinearity.

Another approach to mitigate multicollinearity is to use a correlation matrix as a visual aid to identify highly correlated covariates and perform exclusion based on knowledge about the relevance of the variables [38]. However, a correlation matrix only investigates the relationship between pairs of variables and cannot be used to detect multicollinearity between three or more variables.

A commonly used regularisation technique is L1 regularisation, also known as Lasso regression, which is the method we use in our research. The technique adds a penalty term that is proportional to the absolute value of the coefficients of the model [53]. This technique shrinks some of the coefficients to zero, effectively removing them from the model and providing a form of feature selection.

Equation 3.9 shows Cox regression with L1 regularisation.

$$\arg\max_{\beta} \log PL(\beta) - \lambda \sum_{i=1}^{p} |\beta_i| \tag{3.9}$$

$PL(\beta)$ is the partial likelihood function of the CPH model, $\beta$ are the coefficients for $p$ covariates, and $\lambda \geq 0$ is a hyper-parameter that controls the amount of shrinkage, also known as the penalty term [51].

To obtain the best penalty term $\lambda$ and consequently the optimal model, we regularise the CPH model with different values of $\lambda$. We measure optimality based on two performance metrics: the BIC and C-index. We use cross-validation to evaluate the model for different $\lambda$ parameters based on the BIC and the C-index.

## 3.5 Model Evaluation

In order to select suitable features and find an optimal regularised model, it is important to validate and accurately estimate its performance. In

this section, we discuss the use of cross-validation as a method for model validation and introduce the BIC and the C-index as metrics to evaluate the goodness of the model.

### 3.5.1 Cross-Validation

Cross-validation evaluates the performance of a model and assesses its ability to generalise to new, unseen data [6]. The basic idea of cross-validation is to divide the available data into two sets: a training set and a validation set. The model is trained on the training set, and its performance is evaluated on the validation set. This process is repeated multiple times, with different subsets of the data used for training and validation each time, and the average performance is computed.

The most common form of cross-validation is k-fold cross-validation, where the data is divided into k subsets of approximately equal size. The model is trained on k-1 subsets and evaluated on the remaining subset, with this process repeated k times. The results of each fold are averaged to obtain an estimate of the model's performance.

The number of folds used in cross-validation can affect the performance estimates [12]. A small number of folds, such as 2, may result in a high variance in the performance estimates, as the evaluation highly depends on the subsets chosen. On the other hand, a large number of folds, such as 10 or more, may lead to a high computational cost and may provide little additional benefit in terms of performance estimation. Additionally, too many folds lead to small subsets, which could underfit the model. Using 5-fold cross-validation is most common, as it is a reasonable compromise between the variance and computational cost. It provides a sufficient number of folds to reduce the variance in the performance estimates while still being computationally feasible for many datasets.

When fitting a CPH model using different regularisation values, we use cross-validation for each value to evaluate its performance. We then compare their performance and choose the most optimal one. Their performance can be represented in metrics using the BIC and the C-index.

### 3.5.2 Bayesian Information Criterion

The BIC is a model selection criterion used to compare different models based on their goodness of fit and complexity [46].

BIC is the logarithm of the maximum likelihood of the data under the model, penalised by a term that depends on the number of parameters in the model and the sample size. Mathematically the BIC can be written as:

$$BIC = -2\ln(L) + k\ln(n) \tag{3.10}$$

where $L$ is the maximised value of the likelihood function, $k$ is the number of parameters in the model, and $n$ is the sample size. The penalty term $k\ln(n)$ ensures that models with more parameters get higher BIC values and vice versa. Models with lower BIC values are preferred.

In the context of our cross-validation for different regularisation values, the BIC score is computed for each model on the validation set. The model with the lowest BIC score is considered optimal, as it provides the best trade-off between goodness of fit and complexity.

### 3.5.3 Concordance Index

In survival analysis, the C-index measures the predictive accuracy of a survival model [11]. It measures the proportion of all pairs of subjects whose predicted survival times are correctly ordered among subjects with different survival times. In other words, it represents the probability that a randomly selected pair of subjects will be correctly ordered according to their survival times.

The C-index can be expressed mathematically as:

$$C = \frac{\sum_{i \neq j} 1\{\eta_i < \eta_j\} 1\{T_i > T_j\} d_j}{\sum_{i \neq j} 1\{T_i > T_j\} d_j} \tag{3.11}$$

where $i$ and $j$ are two individuals, $\eta$ is the risk score or hazard, $T$ represents the time-to-event or duration, and $d$ represents if the event occurred or not.

The C-index ranges from 0 to 1, where a value of 0.5 indicates a model that performs no better than chance, and a value of 1 indicates a perfect model. A C-index of 0.7 or higher is generally considered a good survival model performance.

The C-index can be used to compare the performance of different survival models, which in our case is the CPH model with varying values of regularisation. We use the C-index to find the optimal model and assess the impact of different covariates on the model's predictive accuracy.

Using BIC and C-index together can provide a more comprehensive evaluation of the model's performance. BIC is useful for selecting the most provident model, while C-index is useful for assessing the model's predictive accuracy. By considering both measures, we can ensure a simple and accurate model.

## 3.6 Implementation

This section presents the tools we use to implement the methods described above. We use Python [61] as our programming language, as it is easy to use and offers a wide range of powerful libraries and tools ideal for research and development [44]. Specifically, we use the libraries Pandas, NumPy, Lifelines, Matplotlib and Seaborn.

**Pandas** Pandas [59] is a Python library for data manipulation and analysis. It provides easy-to-use data structures and analysis tools for handling structured data, such as tables or spreadsheet-like data. Pandas provides two primary classes for working with data - Series and DataFrame. In our implementation, we use Pandas as part of our

pre-processing to generate the data structures used for survival analysis models, as described in Section 3.3. We create DataFrames with columns for durations, events and additional covariates and rows for each individual or instance.

**NumPy** NumPy [32] is a library that supports large, multi-dimensional arrays and matrices. It also provides a collection of high-level mathematical functions to operate on these arrays. NumPy is a useful package for scientific computing in Python, as it is a powerful and efficient tool for performing numerical computations. We use NumPy for mathematical functions, such as calculating the mean of covariate values or the BIC scores from cross-validation, as described in Sections 3.3 and 3.5.2. We also use the library to generate different arrays, such as a range of $\lambda$ values when looking for the optimal penalty term, as described in Section 3.4. We use NumPy for both pre-processing and in our experiments.

**Lifelines** Lifelines [16] is a library for survival analysis that provides a set of tools for exploring and analysing time-to-event data to make survival analysis more accessible and easier for researchers and data scientists. The library offers different statistical and visualisation tools for performing survival analysis with univariate and multivariate models. It also has support for time-varying covariates. We use Lifelines in our implementation mainly to apply the survival analysis models described in Section 3.2 to our data. We also use the library to structure time-varying covariates, cross-validate our regularised CPH model, calculate C-indexes and plot and visualise analyses and results. The library also provides a function for finding the best univariate parametric model for the data, which fits different parametric models and returns the best performance and accuracy. We use this as our third univariate model after Kaplan-Meier and Weibull to see if other models better fit our data. In our case, this is the Piecewise Exponential model. The library also provides a function for checking if the proportional hazards assumption is met, as discussed in Section 3.2.4. The function performs a statistical test checking time-varying coefficients using four time transformations, which ultimately checks the assumption. We apply this function in our experiments with the CPH model to check that the assumption is not violated. Ultimately, we use Lifelines for pre-processing and the actual survival analysis in our experiments.

**Matplotlib** Matplotlib [34] is a plotting library that provides different tools for creating visualisations and figures, including line plots, scatter plots, bar plots, histograms, and more. It integrates well with our other libraries, making it a valuable tool for our data visualisation and exploration. We use the library to create and save the figures presented in this thesis in combination with the plotting tools from Lifelines. Additionally, we use Matplotlib to analyse our results through visual

interpretations. Ultimately, we use the library for plotting and visualisation in our experiments.

**Seaborn** Another Python library for plotting and visualisation is Seaborn [65], which provides a high-level interface for creating informative and attractive statistical graphics. Seaborn builds on top of Matplotlib and integrates well with Pandas and NumPy, making it easy to use for our implementation. It includes a variety of plot types, including scatter plots, line plots, bar plots, and heat maps, among others. In our research, we use Seaborn to plot our correlation matrix, a type of heatmap, of the covariate values. The correlation matrix is easy to generate and interpret and is part of our process of selecting covariates with low correlation and visualising it in this thesis. Similar to Matplotlib, we use the Seaborn for plotting and visualisation.

We use Python's powerful data analysis and scientific computing libraries for implementation, including Pandas for efficient data pre-processing and data structuring, NumPy for numerical operations, and Lifelines for survival analysis experiments. By using these tools, we can simplify our workflow, reduce implementation time, and focus on interpreting and drawing insights from our results. Additionally, we use Matplotlib and Seaborn to create plots and visualisations that help us better understand our data and results and communicate our findings effectively.

## 3.7 Chapter Summary

This chapter presents our methodology for utilising survival models in our thesis. The models are divided into univariate and multivariate models. The former includes the Kaplan-Meier, Weibull, and Piecewise Exponential models used to validate data distribution for survival analysis. We employ the multivariate CPH model and its extension, the CTV model, for estimating the effects of covariates.

To use these survival analysis models, the data must be structured in a counting process format, including duration intervals for each event occurrence and censoring for individuals who have not experienced the event of interest. We provide examples of the data structures for the univariate models using soccer players and injuries. For the multivariate models, we present several data structure examples incorporating covariates, such as covariate values from the day of the event, averaged values from the duration intervals, and time-varying covariates. We also present a data structure including recurrent events.

We also describe regularisation and how to use it for feature selection and mitigate the issues of overfitting and multicollinearity. Finally, we outline how cross-validation can be employed to evaluate the survival analysis models, explicitly using the BIC and C-index.

This chapter also includes an overview of the different Python libraries and tools we use in our implementation, such as Pandas, NumPy, Lifelines,

Matplotlib and Seaborn. We use these libraries for pre-processing, data structuring, numerical operations, survival analysis, plotting and visualisation.

The following chapter will showcase the process and results of using the methods described in this chapter by dividing the process into four experiments.

# Chapter 4

# Experiments and Results

In the previous chapter, we outlined the various methods employed in this thesis, including survival analysis techniques, models and data structures, and methods for feature selection and model evaluation.

This chapter describes our process and results of using the survival analysis methods described in Chapter 3. We present our data pre-processing steps, followed by these four experiments and their respective outcomes:

**Exp. 1** Validate our dataset for survival analysis using univariate models.

**Exp. 2** Estimate the effect of covariates using the CPH model with covariates from the day of the injury.

**Exp. 3** Estimate the effect of covariates using the CTV model with time-dependent covariates.

**Exp. 4** Estimate the effect of covariates and feature selection using a CPH model with regularisation with averaged covariates.

## 4.1 Pre-Processing

For our experiments, we use the SoccerMon dataset described in Section 2.4 from the years 2020 and 2021. Specifically, we use the subjective data collected from each player on a team, meaning their daily training load and wellness metrics, including injuries. Table 4.1 shows an overview of team injury statistics, including the number of players in total, non-injured players, injured players and the total number of injuries. As discussed

| | **Players** | **Non-injured** | **Injured** | **All injuries** |
|---|---|---|---|---|
| **Team A** | 28 | 15 | 13 | 64 |
| **Team B** | 21 | 17 | 4 | 17 |

Table 4.1: Injury statistics for univariate analysis

in Section 3.1.1, injuries can be divided into first injuries and recurrent injuries. The number of first injuries is equal to the number of injured

players in the table, and the number of recurrent injuries is equal to the total number.

The data is stored in CSV files for each player, so we import and concatenate all the files for each team. Once imported, we pre-process the data to improve its quality and usability with the survival analysis models. This section covers the pre-processing steps we take before using the data in our models, such as recalculating ACWR values, handling missing data, structuring the data in the correct format and handling duplicates for recurrent injuries.

### 4.1.1 Changing ACWR Values

While exploring the training load metrics, we observed inaccuracies in how the ACWR values were calculated. As explained in Section 2.4.1, the ACWR is computed by dividing the acute training load by the chronic training load from the last 42 days. However, the calculation was reversed in the dataset, meaning chronic load was divided by acute load. As a result, we recalculate the ACWR values by dividing the acute training load by the chronic training load from the last 42 days to get the correct measurements.

### 4.1.2 Missing Data

Because missing data is a common occurrence in real-world datasets, it is important to consider how to handle missing values during analysis. Missing data is expected in our case, as we use subjective data reported by players. We observe a significant amount of data missing at the beginning and end of players' observation periods. Various reasons could cause this, such as a player switching teams or a new member joining later. This could also be due to when the match seasons start and end. The teams used in this research participate in the Norwegian soccer league "Toppserien", in which the season of 2020 lasted from July to December and 2021 from May to November [23, 24]. Players might have reported only during match season, resulting in missing data at the start and end of the year.

For univariate models, these missing entries are handled through censoring, where the events are set to false, as described in Section 3.1.2. As we can see from Table 4.1 there are 13 injured players and 15 non-injured players for Team A, meaning there are 13 present events and 15 censored individuals. Because of censoring, we can leave this missing data as is.

Multivariate models, however, are dependent on having present covariate values. To mitigate this issue, we trim off the start and end of each player's dataset where significant missing data occurs, based on present ACWR values as valid first and last entries. We also try using wellness values as the first and last valid entries, but it does not make any difference, so we land on using ACWR for this issue. Table 4.2 shows an overview of injury statistics after trimming off the start and end of the dataset. Team A is left with 6 injured players, as 7 were missing training load and wellness values. Team B is left with 3 injuries, meaning only 1 player was removed.

| | Players | Non-injured | Injured | All injuries |
|---|---|---|---|---|
| **Team A** | 21 | 15 | 6 | 56 |
| **Team B** | 20 | 17 | 3 | 16 |

Table 4.2: Injury statistics for multivariate analysis

As the observation period is between January 2020 and December 2021, cutting off missing entries at the beginning and end results in different observation periods between players. These are still in 2020 and 2021, but some players might have shorter observation periods.

Missing data can also arise from other factors, such as the tedious nature of daily reporting, leading to incomplete or infrequent reporting, or players only reporting on training and match days. This creates missing data points scattered throughout the dataset. We solve this differently for each experiment.

For Experiment 1, we only have to account for missing injuries, as the experiment uses univariate models, which do not handle any covariates. We cannot determine whether missing injury data is due to a player not experiencing injuries or simply forgetting to report them. In either case, we handle these missing events using censoring techniques, as explained in Section 3.1.2. We assign individuals with missing injury data a duration from their first report to their last report, with the event set to false.

In Experiment 2, we apply the multivariate CPH model as discussed in Section 3.2.4, which incorporates covariates from the dataset. The covariates used in this experiment are from the day of the event. However, there are missing data points, partially due to subjective data from two separate reports - one for training load and the other for wellness. We observe more missing values in wellness than in training load. Consequently, there are instances with present ACWR values but missing wellness values. Table 4.3 provides an overview of the number of players, the number of these players missing training load or wellness data, and the number of missing data points and data points in total. If we were to

| | Players | Players missing data | Missing points | All points |
|---|---|---|---|---|
| **Team A** | 21 | 9 | 63 | 168 |
| **Team B** | 20 | 16 | 112 | 160 |

Table 4.3: Missing data statistics for multivariate analysis

remove all players with missing covariate values, we would be left with small datasets. For Team A, we would be left with 12 players and for Team B, only 4 players. Nevertheless, because most of these instances have false events, implying that no injury occurred, they are censored. Thus, we replace these missing points with zeros, as they have no bearing on the analysis. There is one instance in Team B of a reported injury with missing wellness values but present ACWR. Because there is only one instance of this, we replace the missing wellness values with zeros, as it does not significantly impact the outcome.

For Experiment 3, we use the CTV model that accounts for time-varying variables, meaning that all values throughout the duration are used. Naturally, the same issue that occurred in Experiment 2 arises here as well. Table 4.4 shows an overview of missing data in the dataset using all entries throughout the duration intervals. For Team A, only 11 out of

| | **Entries** | **Entries missing data** | **Missing points** | **All points** |
|---|---|---|---|---|
| **Team A** | 6214 | 11 | 77 | 49712 |
| **Team B** | 5101 | 7 | 49 | 40808 |

Table 4.4: Missing data statistics for time-varying analysis

6214 entries are missing training load or wellness data. For Team B, there are only 7 out of 5101 such instances. As before, most of these instances have false events, indicating that the players did not experience any injuries and are therefore censored. There is only one occurrence of an injury with missing covariate values, which is in Team A, which could be due to the player only reporting the injury on that day or reporting the injury later. However, this is a small portion of the data because the model includes all covariate values from whole durations. Consequently, we fill these missing points with zeros since they do not have a significant impact on the outcome.

Lastly, for Experiment 4, we apply a multivariate CPH model similar to Experiment 2, except that we use the averaged values of the covariates instead of their values on the day of the event. Handling missing values here is different from Experiments 2 and 3. Instead of filling in the missing values, we include all available covariate values from each duration interval and calculate their mean. As we can see in Table 4.4 when including all covariate values from durations, there are few missing data points compared to the total amount. Therefore, we still have many values to calculate averages from. This approach allows us to avoid handling missing values explicitly, as the calculation uses only the available data points.

Our first research question in Section 1.2 ask how missing data should be handled. In this section, we answer this question. Missing data is handled differently for each experiment. Ultimately, when applying univariate survival analysis models in Experiment 1, missing data is handled through censoring. For multivariate models dependent on covariate values, the beginning and end of the dataset are removed based on missing ACWR values. For Experiments 2 and 3 using day-of-the-event covariate values, missing data points are filled with zeros, as most of them are from non-injured players and do not significantly affect the outcome. For Experiment 4, using averaged covariate values, missing data points are left as is because the calculation uses only the available data points.

### 4.1.3 Data Structuring

As explained in Section 3.3, the data in survival analysis is organised using a counting process format. We follow the structures presented there using data frames, with columns for durations and events and the corresponding covariate values for multivariate models. To accommodate first and recurrent injuries, we generate distinct structures for each. As each experiment requires a unique data structure, we present the specific format with the corresponding experiment in Sections 4.2, 4.3, 4.4 and 4.5.

### 4.1.4 Duplicate Injuries

Experiments 2 and 4 include recurrent injuries. This poses a challenge, as players may report the same injury over multiple days, making it difficult to distinguish between a new injury and the same injury. To address this and avoid duplicates, we set a limit of 5 days between injuries. However, this approach may also exclude separate injuries that occur within the same 5-day window. Although this is a potential limitation, we have accepted this trade-off to avoid counting the same injury multiple times.

## 4.2 Experiment 1 - Univariate Models

In our first experiment, we assess the suitability of our dataset for survival analysis by using univariate survival analysis models. Specifically, we employ the Kaplan-Meier, Weibull, and Piecewise Exponential models described in Section 3.2.

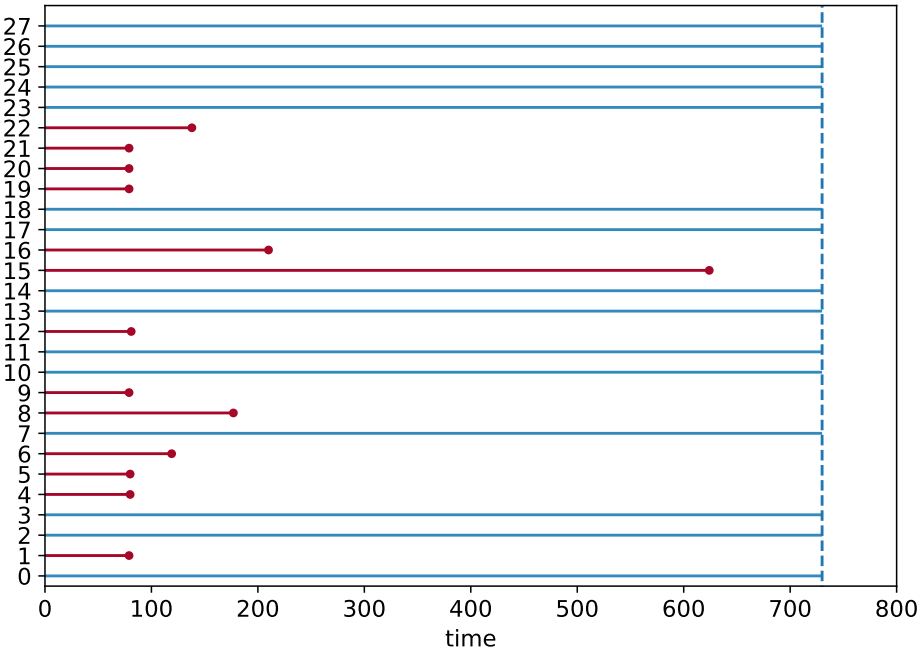
### 4.2.1 Setup

As mentioned in Section 4.1.3, we structure the data differently for each experiment. For Experiment 1, we generate a data frame with a duration column and an event column, where each row is a first injury instance from a player or a censored instance if a player has had no injuries. We calculate the duration as the number of days from the first report to the day of the first injury. For players with no injuries, we calculate the duration as the number of days between the first report and the last report in their observation period.

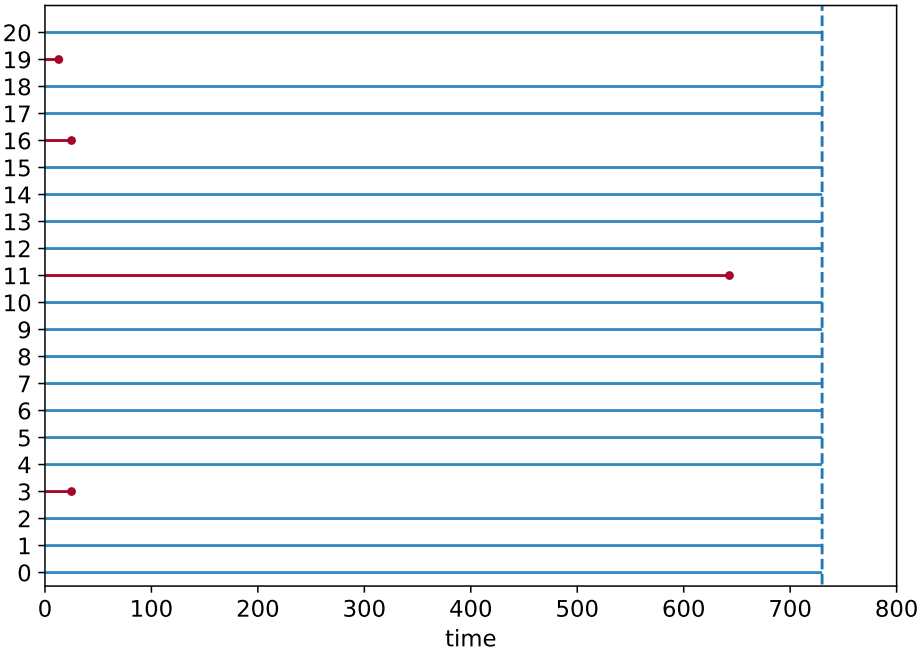
We do not include recurrent injuries in this experiment, as the univariate survival analysis models would treat these as if they are from separate individuals and would not provide results accurate to the true events.

Figure 4.1 illustrates the players and their first injury events over time, including censoring. The red lines represent players with injuries during the observation period, and the blue lines represent the censored players.

After pre-processing and organising the data, we compare the injury distribution in frequency over time with the survival functions derived from the models. The number of injuries that occur at a specific point in time provides insight into the likelihood of experiencing an injury at that




(a) Team A




(b) Team B

Figure 4.1: Injury events and censoring for univariate analysis

point. Consequently, we can leverage the survival functions derived from the survival analysis to determine if the injury distribution aligns with our expectations.

To compare the Kaplan-Meier, Weibull and Piecewise Exponential models with the injury distribution, we generate a histogram of the distribution and plot the survival function of each model. This approach provides a clear visual representation of the relationship between the models and the injury distribution. We perform this analysis for both teams, Team A and Team B, in order to assess if the injury data follows a distribution suited for survival analysis models

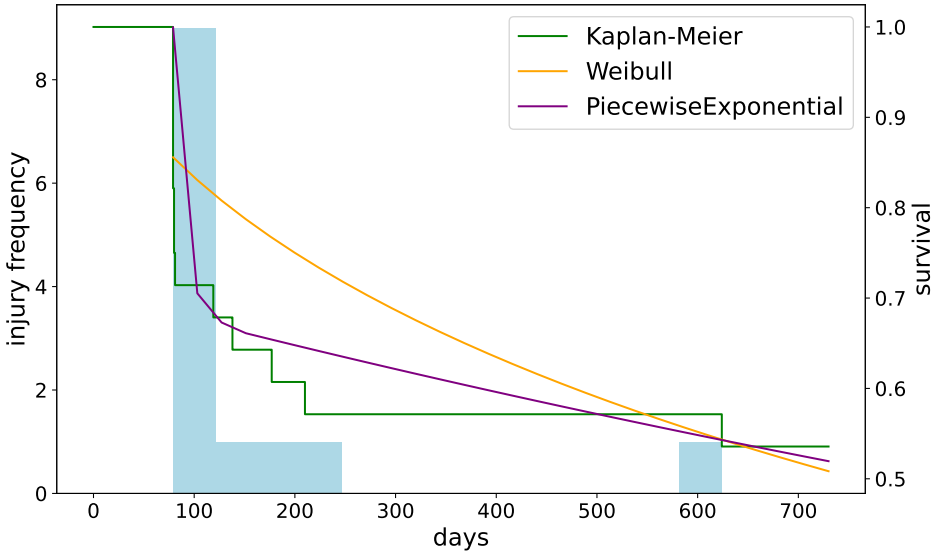
### 4.2.2 Results

The results from using first injuries from both teams are shown in Figure 4.2. The injury frequency axis is related to the histogram, and the survival axis is related to the survival functions of the models. We plot this over the time interval for which the data has been collected, which is 730 days, during which the survival probability decreases from an initial value of 1.0. Both teams have few first injuries, as shown in the histograms, and most of them occur at the beginning of the observation period.

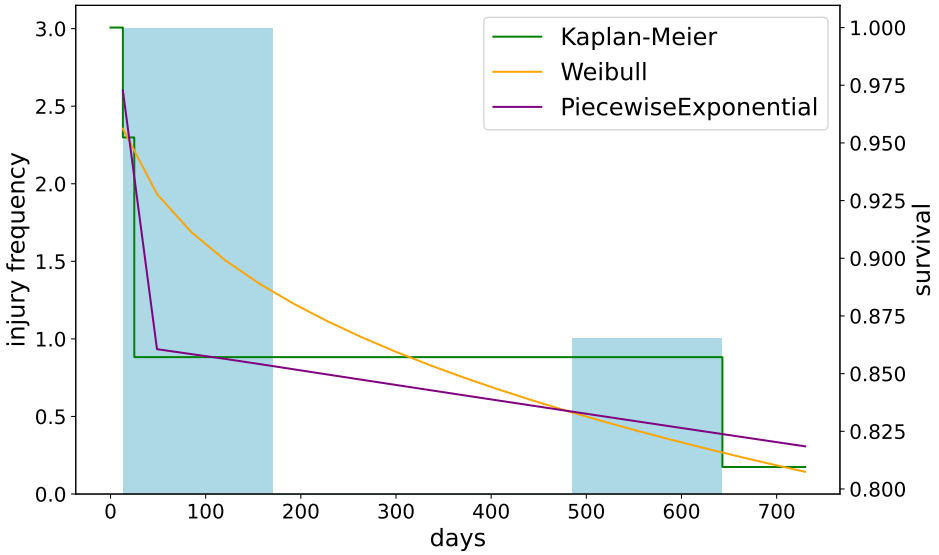
However, the slopes of the survival functions do match the injury distribution, as they both have a steep decline where most of the injuries occur. For Team A, this is around day 100, as most first injuries occur here. This might be due to late reporting, as many first injuries occurred on the same day. The Kaplan-Meier corresponds to each injury occurrence as it steps down on the survival probability for each instance. From approximately day 200 to day 600, the Kaplan-Meier indicates a 60% chance of not having an injury, meaning that the probability decreased by 40% in the first 200 days. The survival function of the Piecewise Exponential model has a more accurate slope than Weibulls, as it has a considerable decline around day 100 and flattens out as fewer injuries occur.

For Team B, with even fewer first-injury occurrences than Team A, the Kaplan-Meier suggests a survival probability of approximately 85% from around day 25 to day 650, which is high for such a large time interval. The slopes from the Weibull and Piecewise Exponential models are similar to the survival functions in Team A, and again the Piecewise Exponential model is more accurate than Weibull. Due to the size of Team B's dataset and these results giving such a high survival probability, we keep in mind that Team B's dataset might be more susceptible to errors or outliers that could affect the results and proceed by interpreting results from Team B with caution.

We conclude that survival analysis is possible for our dataset from these findings. We also answer one of our research questions from Section 1.2 regarding using univariate models for validating our dataset by proving that the injury distribution of our data matches the survival functions. Therefore, we can use univariate survival analysis models to validate our injury data for survival analysis.




(a) Team A




(b) Team B

Figure 4.2: Univariate analysis and injury distribution

## 4.3 Experiment 2 - Cox Proportional Hazards Model

After validating our dataset for survival analysis in Experiment 1, we proceed to identify the effects of the covariates using the CPH model.

### 4.3.1 Setup

The data is structured as described in Section 3.3.2 for the CPH model using first injuries. We generate a data frame with durations, events and covariate values from the day of the event. We include the training load and wellness values described in Section 2.3.1 as covariates. Ultimately, each row in the data frame includes a duration, an event, and a value for each covariate on the day of the event. If the instance is censored, the covariate values are from the day of the last report.

Figure 4.3 illustrates the players and their first injury events over time, including censoring, after trimming off missing ACWR values for multivariate analysis, as described in Section 4.1.2. The red lines represent players with injuries during the observation period, and the blue lines represent the censored players. As we can see, the number of injuries is reduced, especially for Team A. Observation periods are also reduced and vary from player to player. This reflects the trimming of the beginning and end of the players' datasets, as some only have reported for a short period, and others have reported for almost the whole observation period.

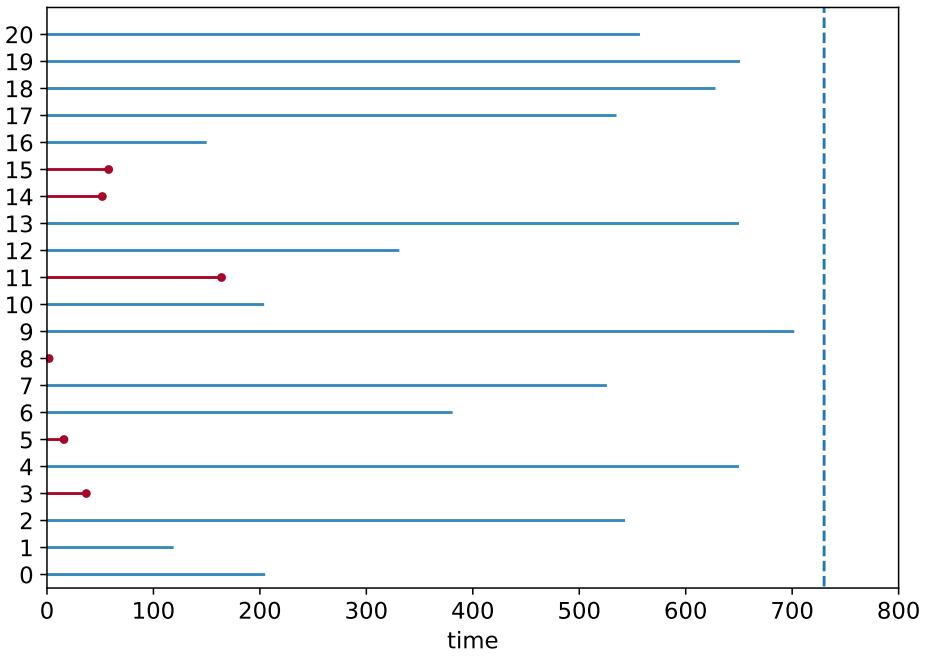
We introduce recurrent events as described in Section 3.1.1 and generate a data frame following the data structure presented in Section 3.3.2. Figure 4.4 shows the players and their recurrent injury events over time, including censoring. We observe that most injured players have experienced more than one injury during the observation period. Therefore, including recurrent injuries in our analysis provides a more realistic depiction of the data, as injuries are often not terminal events for soccer players.

When considering only the first injury of each player, the dataset is smaller and might not fully represent the complexity of injury patterns in the team. However, including recurrent injuries allows for a larger dataset, including more observations and data, providing more statistical power for the analysis. This, in turn, can lead to more accurate estimates of the effects of covariates on injury risk.

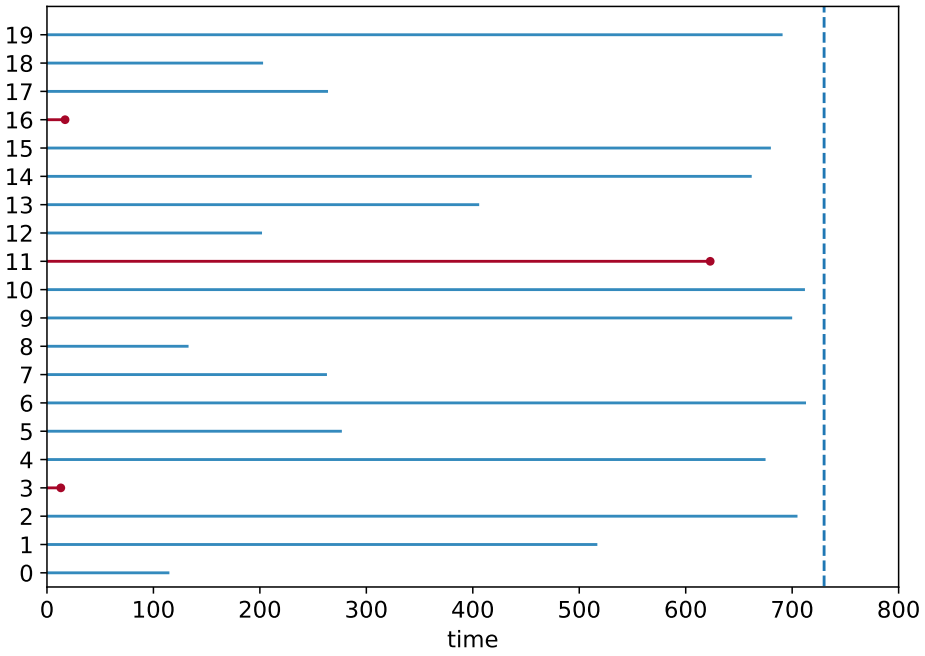
Additionally, recurrent events have the benefit of providing time-varying covariates, as we can analyse the values from the different duration intervals of the same individual. The CPH model assumes that events are independent, but we want to capture how prior events may affect other events. Therefore, we include a new binary variable, prior injury, stating whether a player has had an injury previously in the observation period.

We have 15 training load and wellness variables available, and including them could potentially lead to issues such as overfitting and multicollinearity, as discussed in Section 3.4. Therefore, we need to select which covariates to include in our analysis.

As discussed in Section 2.6, ACWR is a widely used measurement for assessing injury risk and has proven useful in previous research [64].

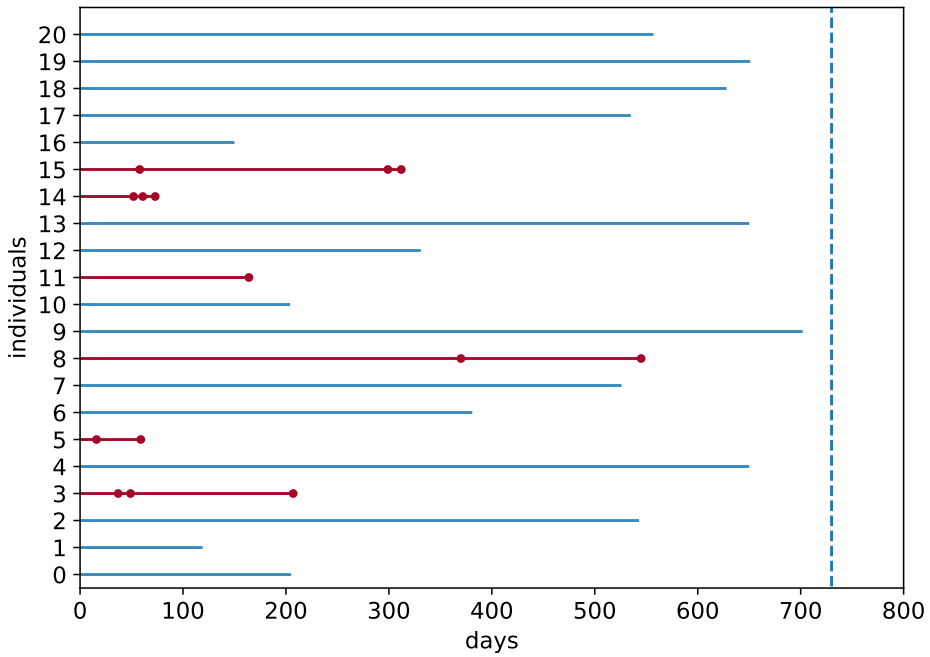



(a) Team A

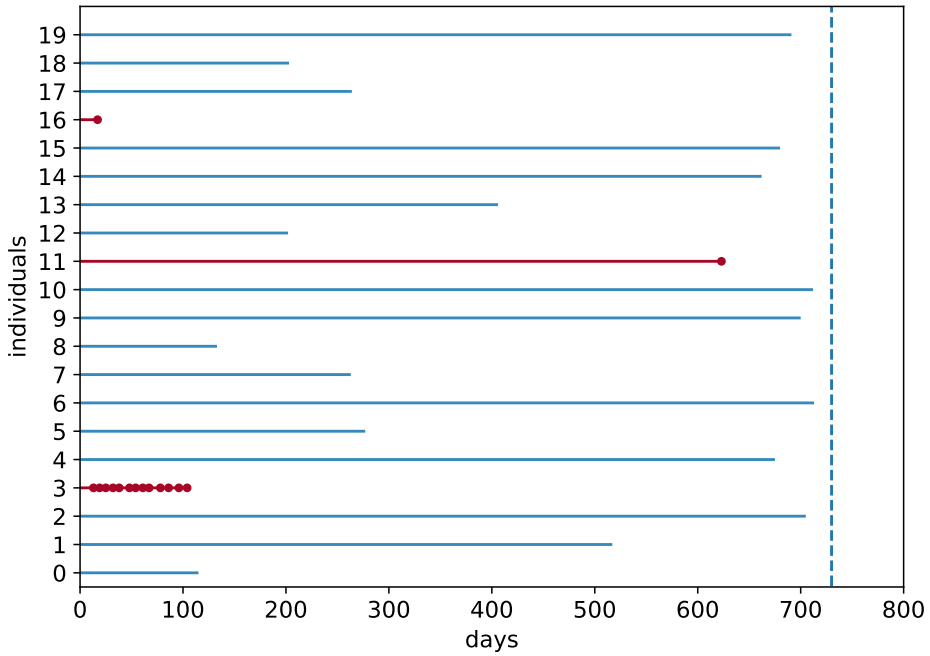



(b) Team B

Figure 4.3: Injury events and censoring for multivariate analysis using first injuries




(a) Team A




(b) Team B

Figure 4.4: Injury events and censoring for multivariate analysis using recurrent injuries

Therefore, we decide to include this training load metric as one of our covariates. Additionally, there have been discussions and concerns regarding using Gabbett's injury risk model based on ACWR, stating that there is a "sweet spot" for ACWR values between 0.8 and 1.3 where injury risk is lowest [25]. The model suggests that significantly low and high ACWR values increase the risk of injury. Therefore, another reason to include ACWR as one of our covariates is that we can investigate the accuracy of this injury risk model.

For selecting the other covariates for our analysis, we use a correlation matrix as a visual aid to identify variables with high correlation. Excluding these variables helps us reduce the number of covariates while ensuring they are not highly correlated. Our correlation matrix is shown in Figure 4.5, where darker colours indicate a high correlation between the observed covariates and lighter colours indicate a low correlation. We observe that

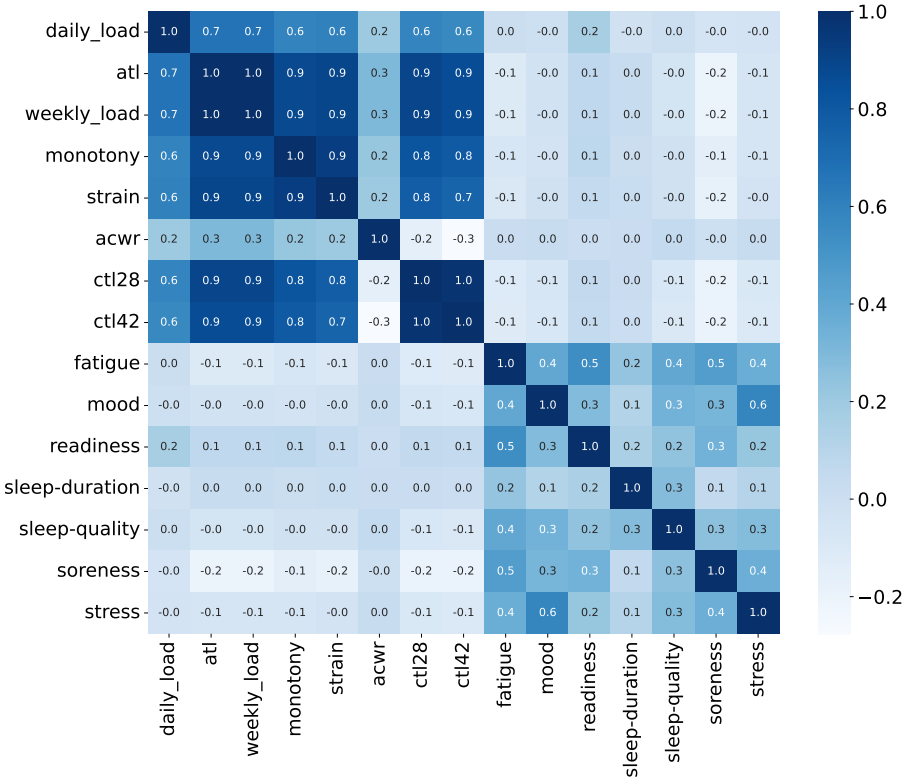



Figure 4.5: Correlation matrix of training load and wellness variables

some of the training load variables are strongly correlated due to their interdependence when they are calculated. For instance, weekly load and Acute Training Load (ATL) have a correlation value of 1.0, meaning 100% correlation. This is because the weekly load is calculated as the sum of sRPE over the last 7 days, and ATL is the average sRPE over the last 7 days [42]. It is therefore expected that they will change similarly. However, we observe that ACWR has a low correlation with the other training load metrics, supporting it as a suitable candidate for our analysis.

Our correlation analysis shows that the wellness variables in our dataset have a relatively low correlation, with the highest correlation being between mood and stress, with a coefficient of 0.6. The correlation between wellness and training load metrics is also relatively low.

Based on these findings, we conclude that ACWR, fatigue, mood, readiness, sleep duration, sleep quality, soreness, and stress are suitable covariates for our analysis. We also include our binary variable, prior injury, for our recurrent events analysis.

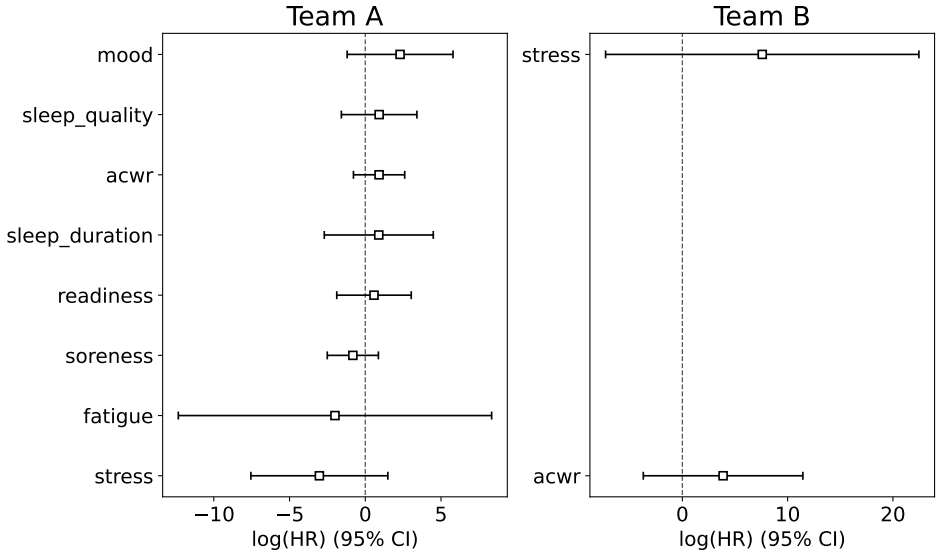
After structuring our data for first and recurrent injuries and selecting appropriate covariates for both teams, we begin our experiment using the CPH model to investigate the effects of the chosen covariates. We apply the model to both data structures and both teams for comparison. As described in Section 3.2.4, the proportional hazard assumption must be met for the CPH model. Therefore, we check the assumption for our model using the provided function in Lifelines [16], as described in Section 3.6.

### 4.3.2 Results

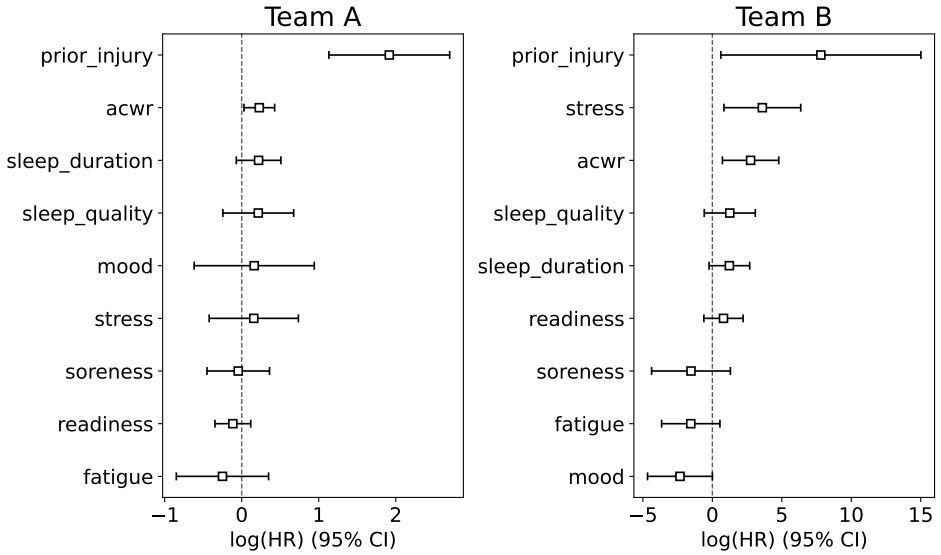
When applying the CPH model to first injuries for Team B, we encounter multicollinearity issues. As described in Section 3.4, this is the issue of multiple independent covariates being correlated in the dataset. Even though we exclude highly correlated covariates using our correlation matrix, there could still be multicollinearity between three or more covariates, which is not visible in the matrix. In order to avoid multicollinearity, we perform an iterative process to identify the best combination of covariates for Team B. We found that the most suitable covariates for this dataset are ACWR and stress, which enables us to avoid multicollinearity issues. Therefore, we only use these two covariates for Team B when examining first injuries.

The results of the CPH model using first and recurrent injuries from both teams are presented in Figure 4.6, displaying the coefficients of each covariate and their corresponding lower and upper bound of the 95% confidence interval. As explained in Section 3.2.4, the confidence interval is the mean of the coefficients plus minus its variations and reflects the certainty of the coefficients[7]. The coefficients indicate how much linear impact a covariate has on the survival outcome. A coefficient higher than zero indicates that increasing the corresponding covariate would increase the hazard. Conversely, a lower than zero coefficient indicates that increasing the corresponding covariate would decrease the hazard. Covariates with coefficients close to zero are considered to have little impact on the outcome.

For Team A, the covariates with the most effect on hazard for first injuries are mood, fatigue and stress, with coefficients around 2.3, -2.0 and -3.0, respectively. The results suggest that increasing mood increases injury risk, and increasing fatigue and stress decreases injury risk. In other words, the better the mood, the more likely a player is to encounter an injury, and the higher the fatigue and stress, the less likely a player is to encounter an injury. Fatigue is measured on a scale of 1 to 5, with 1 indicating feeling very tired and 5 indicating feeling very fresh. These results suggest




(a) First injuries




(b) Recurrent injuries

Figure 4.6: CPH analysis using day-of-the-event covariate values

that the more fresh a player feels, the more they lower the risk of injury. Additionally, the coefficient of fatigue for Team A has a large confidence interval, and as discussed in Section 3.2.4, this indicates that there is a high level of uncertainty in the estimate of its true value, likely due to the small size of the dataset.

For Team B, the results for first injuries suggest that increasing stress levels and ACWR scores increase the risk of injury. As with fatigue for Team A, the coefficient of stress for Team B has a large confidence interval, indicating that the estimated coefficient may be unreliable and that the sample size is relatively small.

For recurrent injuries, the most informative covariates for Team A are prior injury, ACWR and fatigue, with coefficients around 1.9, 0.23 and -0.25, indicating that having prior injuries and higher ACWR increases injury risk and reducing fatigue decreases injury risk. The coefficients for recurrent events are much closer to zero than for first injuries, meaning they have less impact on the outcome. However, the confidence interval for fatigue is much smaller than for first injuries. Moreover, the prior injury covariate provides valuable insight into how much impact prior injuries have on injury risk.

For Team B, all the chosen covariates are included, as there were no issues with multicollinearity. This indicates that the issue of multicollinearity could also occur due to the small sample size, as using more data from Team B avoids this issue. The most informative covariates are prior injury, stress and ACWR, with coefficients around 7.8, 3.6 and 2.8, respectively. These results are similar to Team B's results for first injuries, which indicates that higher stress levels and ACWR increases injury risk. Prior injury also greatly impacts injury risk, as for Team A. However, the prior injury coefficient has a large confidence interval, indicating uncertainty with the estimate of the coefficient.

Ultimately, using first injuries results in more coefficient variance but wider confidence intervals and reduces the possible covariates, likely due to the small dataset. Including recurrent injuries provides a larger dataset resulting in more certainty in the coefficients and allows for the inclusion of more covariates.

We also checked the proportional hazard assumption for our CPH model and found that the assumption is met for both structures and teams.

By revisiting our research questions from Section 1.2, we can answer some of them from what we learned in this experiment. Regarding the question on dataset size affecting the results, we have found that a smaller data set, such as for Team B or when using first injuries only, does result in more unreliable estimates. When including more data, such as for Team A or recurrent injuries, we found that the estimates of the coefficients show more certainty and larger datasets allow for more covariates to be analysed.

The question regarding if the number of covariates affects the results can also be answered, as we found that using too many covariates in our analysis with a small dataset results in issues with multicollinearity. When reducing the number of covariates, this issue is avoided.

We can also answer the question regarding using first or recurrent injuries. We have found that recurrent injuries can provide a realistic picture of soccer injuries, a larger dataset, an additional covariate, prior injury, and reflect time-varying covariates. First injuries might reflect more immediate factors affecting injuries, while recurrent injuries might reflect more long-term effects. Therefore, we conclude that recurrent injuries are more appropriate for our soccer-related research.

In this experiment, we use the covariate values from the day of injury or the last observation day if censored. However, the effect of these covariates may not be immediate but rather occur in the period before the injury. Hence, we aim to investigate the covariates over the entire duration

interval and assess how their changes over time could impact the survival outcome. We use the CTV model in our next experiment to tackle this issue.

## 4.4 Experiment 3 - Cox Time-Varying Model

In the previous experiment, we analysed the significance of covariates using the CPH model, with values from the day of the event. In this experiment, we consider values from the entire duration interval. We do this to capture their time-varying nature and how they may impact the survival outcome.

### 4.4.1 Setup

We structure our data following the approach described in Section 3.3.2 for the CTV model. Each instance corresponds to an observed state change in the duration interval, where days are included only if there is a change in the injury event or a change in training load or wellness. For instance, if there are no changes for several days, these days are excluded. However, they are included in the calculated start and stop points. The start point represents the last state change, and the stop point represents the current state change. The corresponding covariate values are from the day of the stop point. Because we have many covariates, every report has state changes, meaning that start and stop only have one day apart.

In this experiment, we are limited to using only first injuries as our CTV model does not support recurrent events. The data structure we create for the model is extensive, with 6214 rows and 49 712 data points for Team A, as shown in Table 4.4. This is due to including all state changes, which are all daily reports from the players. Building this structure is a complex and time-consuming process. If we were to include recurrent injuries, assuming the model supported it, the process would become even more intricate and prone to errors. Therefore, we decide to use only the first injuries in this experiment. We utilise the same covariates as in Experiment 2, including ACWR, fatigue, mood, readiness, sleep duration, sleep quality, stress, and soreness.

As explained in Section 3.2.5, the CTV model violates the proportional hazards assumption, as it includes time-varying covariates. When the values of these covariates change at different rates for different individuals, the proportional hazards assumption is not fulfilled. This is because the effect of the covariate on the hazard rate may change over time, making it non-proportional. Therefore, we do not have to check that the assumption is met for this experiment.

### 4.4.2 Results

The results from applying the CTV model to first injuries only for both teams are shown in Figure 4.7. The coefficients obtained for both teams are almost equal to zero, indicating that the covariates exhibit very low

variance. This finding suggests that the chosen covariates have little to no influence on the injury outcome over time. However, we observe that the confidence intervals of the coefficients are pretty wide for both teams. Moreover, as explained in Section 3.2.4, this suggests there is uncertainty in the estimated coefficients, which could occur due to a small sample size. This also limits the model's ability to detect significant associations between the covariates and injury risk. Additionally, as described in 3.2.5, the CTV model violates the proportional hazards assumption, which can lead to unreliable results.

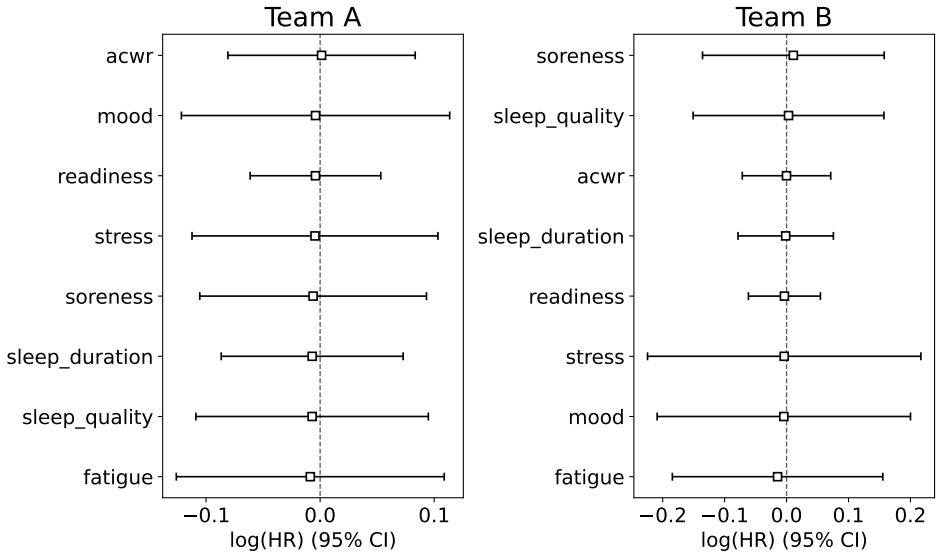



Figure 4.7: CTV analysis using first injuries

In Experiment 2, we found that reducing the number of covariates used in the analysis can result in more coefficient variance. Therefore, we apply the same idea here, using the covariates ACWR and stress. Figure 4.8 shows the results of reducing the covariates for the CTV model. The coefficients still have low variance and wide confidence intervals, supporting the suggestion that the dataset is too small.

These findings support our answer to the research question regarding dataset size affecting the results, as we have found that using a small sample size with the CTV model results in no significance in the effect of the covariates. Including recurrent injuries would enlarge the dataset and could provide more precise results. However, we cannot conclude this with certainty, as we cannot compare the use of the model with a larger dataset, and there could be other reasons for these results.

Focusing on time-dependent variables and their changes assumes that changes in these variables over time are associated with changes in injury risk. However, it is possible that the covariates are not sensitive enough to capture changes that are relevant to injury risk over time. It is also possible that injury risk is more strongly influenced by other factors not captured by the time-dependent variables we included in our model. Therefore, the

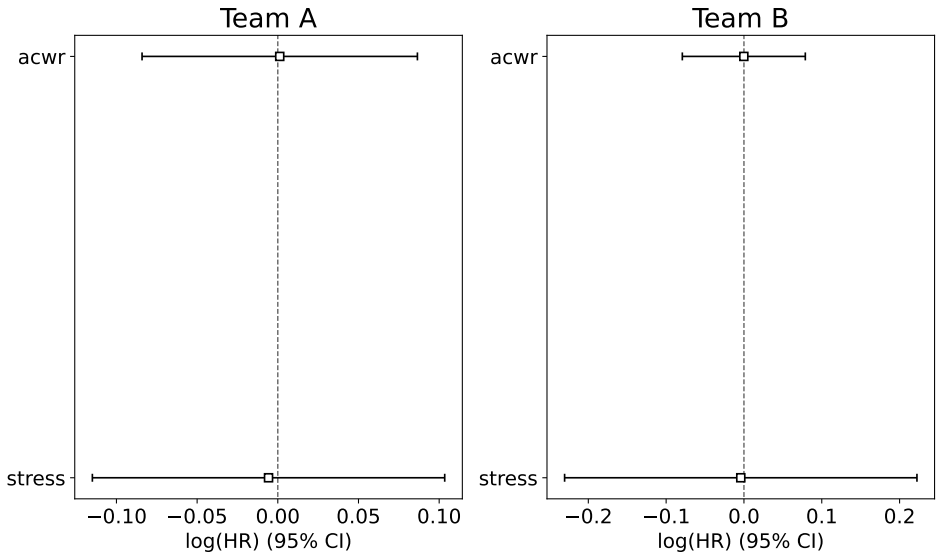



Figure 4.8: CTV analysis using first injuries and a reduced number of covariates

lack of significant associations between the covariates and injury risk may indicate that time-dependent variables are not the primary drivers of injury risk in our case.

Although the CTV model does not yield informative results for our dataset, we still want to capture the potential impact of covariates on the outcome over time. Experiment 2 did suggest that our covariates impact injury risk, so we aim to use the traditional CPH model again but include changes in the covariates using other alternatives. To achieve this, we investigate other options in our next experiment.

## 4.5 Experiment 4 - Cox Proportional Hazards Model With Regularisation

In this experiment, we explore alternative methods to incorporate time-varying covariates based on the findings from the CTV model in Experiment 3. We revisit the CPH model but format the data differently . We also introduce regularisation for feature selection to extract the most significant injury factors.

### 4.5.1 Setup

We follow the data structure for the CPH model described in Section 3.3.2, generating a data frame with columns for durations, events and covariates. We extract all the covariate values from each duration interval and calculate the mean for each covariate. This approach provides a more comprehensive measure of the covariate values across the entire interval,

as opposed to only using the values from the day of the event, which was the case in Experiment 2.

However, calculating the mean across duration intervals may not capture the changes in time-varying covariates as accurately as the CTV model does. To address this, we use recurrent events as we did in Experiment 2. By including all the injuries and calculating the mean of each covariate interval, we capture changes in the covariates over time. We include the prior injury covariate and capture how previous injuries affect other injuries.

We perform feature selection to extract the covariates with the most significant impact on the outcome and eliminate the other covariates. As explained in Section 3.4, regularisation can achieve this goal while preventing overfitting and multicollinearity. We apply regression with L1 regularisation because it can eliminate covariates with low or no effect by setting their coefficients to zero. This allows us to extract the covariates that impact injury risk most.

Regularisation requires adding a penalty term to the model's cost function, which can be adjusted to control the degree of regularisation applied to the model [50]. In order to find the most optimal penalty term which will yield the best results, we fit the CPH model with different values of $\lambda$ representing the penalty term. Using a $\lambda$ value of zero would mean that the penalty term has no effect. Therefore we start with a low number, precisely 0.01, to see how a small penalty term affects the outcome. By testing a range of values for $\lambda$, we found that all coefficients are set to zero or eliminated at around $\lambda = 0.7$, as seen in Figure 4.11. Therefore, we set an upper limit of 0.9. From 0.01 to 0.9, we specify 20 steps within that range. We choose 20 as this allows us to regularise with many different values but does not take too much time in terms of complexity, taking into account the trade-off between exploring a range of values and computational resources.

We use cross-validation to evaluate which value of $\lambda$ results in the best model, as described in Section 3.5.1. Specifically, we use 5-fold cross-validation. The number of folds used in cross-validation can affect the performance estimates. While a small number of folds may result in a high variance in the performance estimates, a large number of folds may lead to a high computational cost [12]. Therefore, we choose 5-fold cross-validation as a reasonable compromise between the variance and computational cost. The data is split into 5 equal-sized subsets. The model is then trained on 4 of these subsets and evaluated on the remaining subset. We repeat the process 5 times using each subset once for validation. The results from each fold are then averaged to give an overall estimate of the model's performance. The estimates we use for performance are the BIC and C-index, as described in Sections 3.5.2 and 3.5.3. BIC balances the model's goodness of fit with the number of covariates in the model, using a penalty term for the number of covariates [46]. In other words, it penalises models with more covariates. We want a $\lambda$ value that results in a low BIC, meaning a model that does not include too many covariates. The C-index is a measure of discrimination that quantifies the ability of the model to correctly rank the observed survival times of pairs of individuals [11]. In

other words, how well a model predicts which of two events will happen first. The C-index ranges from 0 to 1, with a value of 0.5 indicating random prediction and a value of 1 indicating perfect prediction. The closer the C-index is to 1, the better the model performance in terms of discriminating between different survival outcomes. Ultimately, we want a value of $\lambda$ resulting in a model with a high C-index and a low BIC value.

Once we have determined the optimal value of $\lambda$ for our L1-penalized CPH model and obtained a set of covariates with non-zero coefficients, we further refine our selection by filtering out covariates whose coefficients are not significantly different from zero. We use a threshold of 0.1 for the absolute difference between the coefficient and zero to do this. As we explained in Section 3.2.4, the coefficient's magnitude represents the corresponding covariate's impact on the hazard rate. For example, a positive coefficient of 0.1 indicates that a unit increase in the covariate results in a 10% increase in the hazard rate. In contrast, a negative coefficient of -0.1 suggests a 10% decrease in the hazard rate. By requiring a minimum difference of 0.1, we aim to select covariates with a significant impact on the hazard rate, corresponding to at least a 10% change per unit change in the covariate.

As described in Section 3.2.4, the proportional hazard assumption must be met for our CPH model. Therefore, we also check that the assumption is met for the optimal model using the provided function from Lifelines [16], as described in Section 3.6. When we have extracted our covariates, we assess their effects by applying different values to their coefficients. Additionally, we determine their partial effect on survival outcomes using varying values for each covariate.

### 4.5.2 Results

The results from applying the CPH model to first injuries using all our chosen covariates and averaging their values are presented in Figure 4.9. This results in almost no variance in the coefficients and considerable confidence intervals, indicating unreliable estimates, likely due to first injuries being a small dataset. Because of these findings, we reduce the number of covariates until there is more certainty in the coefficients, which results in the covariates ACWR, sleep duration, readiness and fatigue.

The results from using averaged covariates for both first and recurrent injuries, with a reduced number of covariates for first injuries, are presented in Figure 4.10.

The coefficients for ACWR and fatigue show the most significance for first injuries in Team A. However, their confidence intervals are quite large. For Team B, all of the coefficients have large confidence intervals. This indicates that although we reduce the number of covariates, we still get high uncertainty in the coefficients.

The most informative covariates for recurrent injuries in Team A are prior injury, ACWR and sleep quality, with coefficients at 2.4, 0.6 and -1.1. For Team B, these are prior injury, sleep quality and fatigue, with coefficients at 9.4, 5.0 and -2.4. Based on these results, better sleep quality

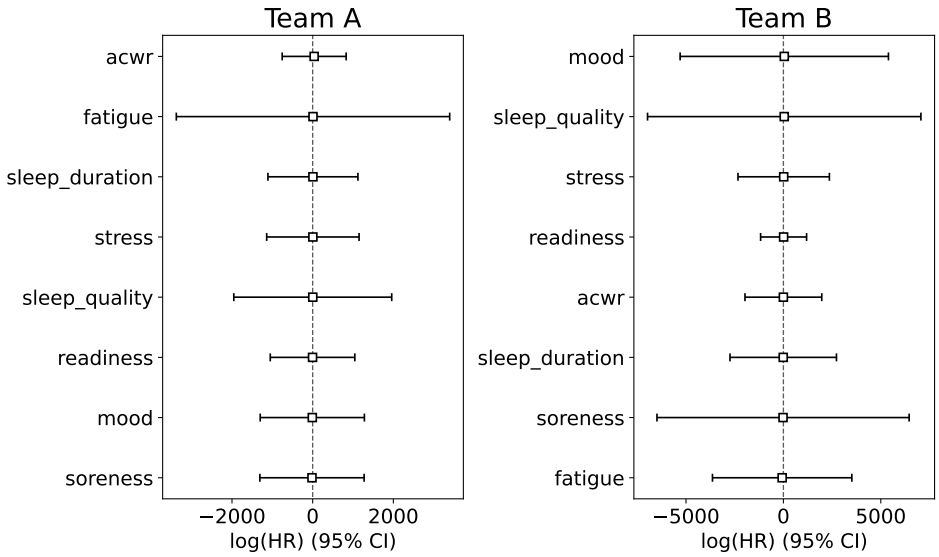



Figure 4.9: CHP analysis using first injuries and averaged covariates

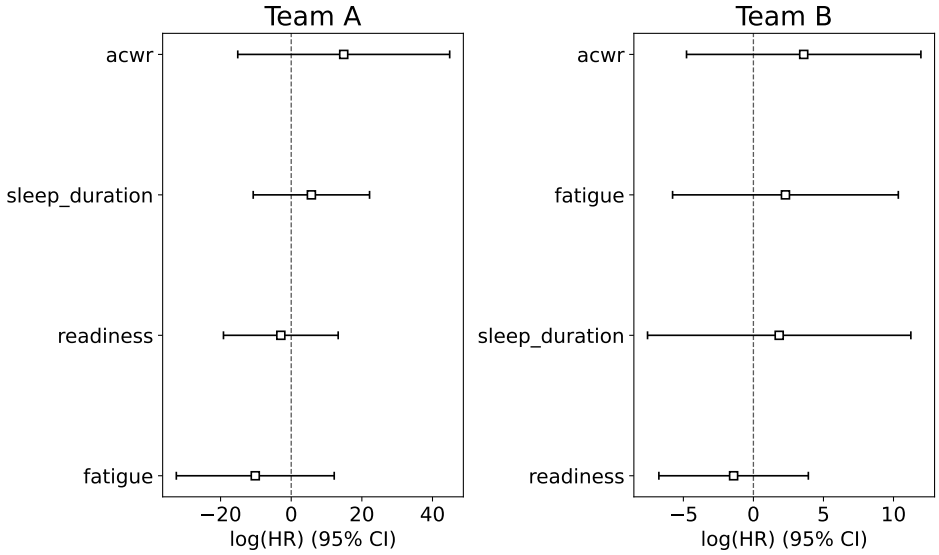
is considered a factor that reduces the risk of injury for Team A, but it increases the risk of injury for Team B.

We observe that using recurrent injuries results in more variance in the coefficients and smaller confidence intervals and allows for more covariates. More covariates allow us to select the most relevant predictors, leading to more accurate and meaningful results. Therefore, recurrent injuries allow a more comprehensive analysis of injury risk factors.

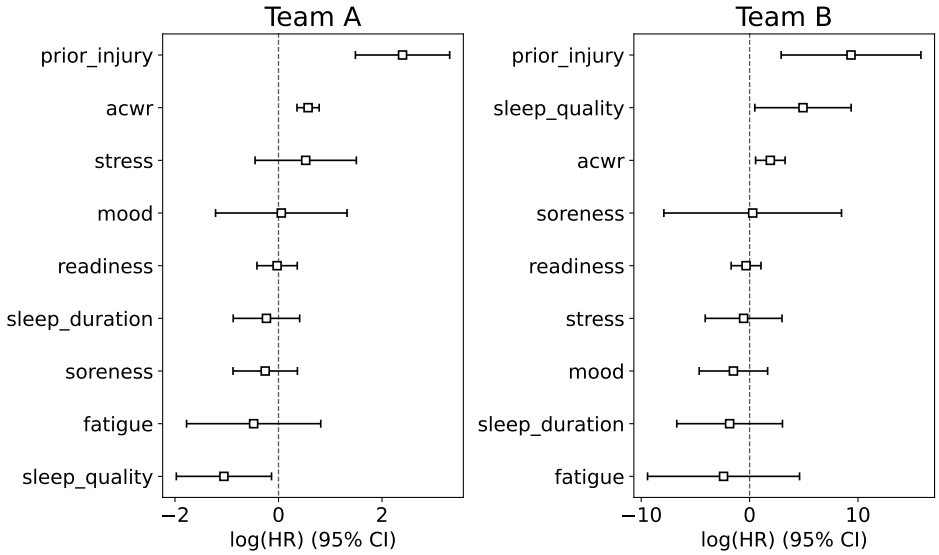
Based on the previous experiments and this experiment, the results from Team B appear to be anomalous, likely due to their relatively small dataset size. We also found in Experiment 2 that the small sample size might be the reason for issues with multicollinearity. Therefore, to ensure the reliability and consistency of our analysis, we continue using results from Team A exclusively. Figures with results from Team B are provided in the appendix in Section A.1.

We use only recurrent injuries from Team A due to the previous results and fit the CPH model with 20 different penalising values from 0.01 to 0.9. The results are presented in Figure 4.11. The x-axis represents the different values of $\lambda$ for regularisation, and the y-axis represents the coefficients of the covariates. The covariates quickly reduced to zero and eliminated by the L1 regression are mood, sleep duration, soreness and stress, meaning they have low or no significance on the outcome. The covariates kept longer by the regularisation, thus having more impact, are prior injury, readiness, ACWR, fatigue and sleep quality. All the covariates have converged to zero at around $\lambda = 0.7$.

We cross-validate and extract the corresponding BIC values and C-index for each CPH model with a different $\lambda$ value. These are presented in Figure 4.12. The x-axis represents the different $\lambda$ values, the left y-axis represents the average C-index, and the right y-axis represents the average




(a) First injuries




(b) Recurrent injuries

Figure 4.10: CPH analysis using averaged covariates

BIC values. The $\lambda$ value with the highest C-index and lowest BIC value is quite low, precisely 0.104.

Based on the $\lambda$ value giving high C-index and low BIC, we fit our optimal CPH model with the $\lambda$ value as the penalty term and the chosen covariates. The results are displayed in Figure 4.13. The chosen covariates are prior injury, ACWR, fatigue and sleep quality. We also found that the proportional hazard assumption is met for the optimal model.

The risk function for each covariate based on the coefficients from the optimal CPH model is displayed in Figure 4.14. By applying a range of values to the covariate's coefficients, we see how they change and how

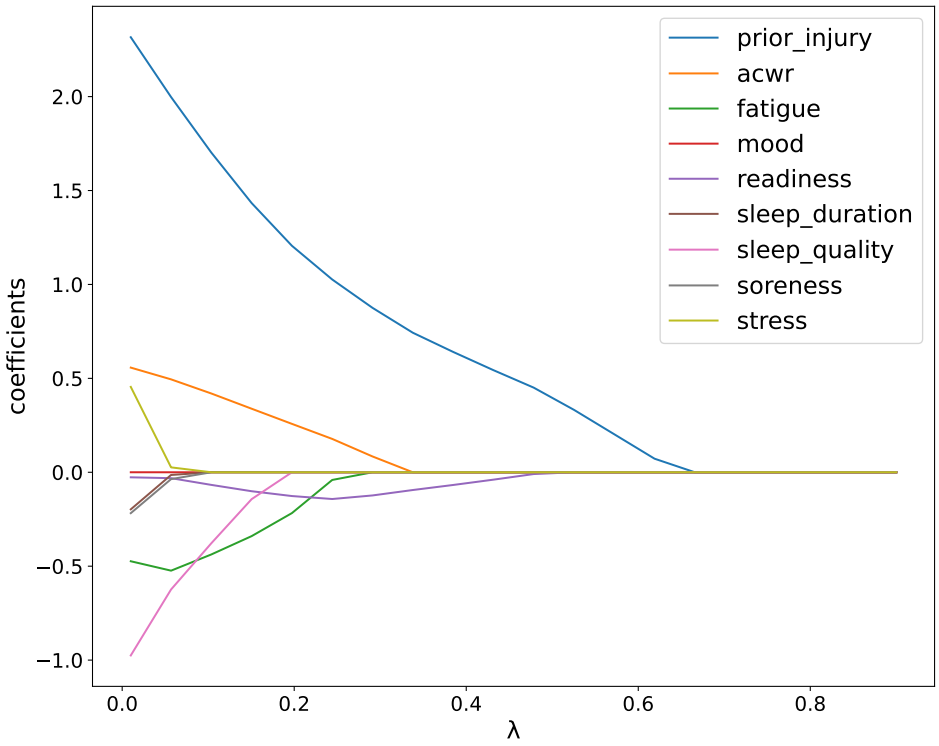



Figure 4.11: CPH analysis with L1 regularisation using recurrent injuries from Team A

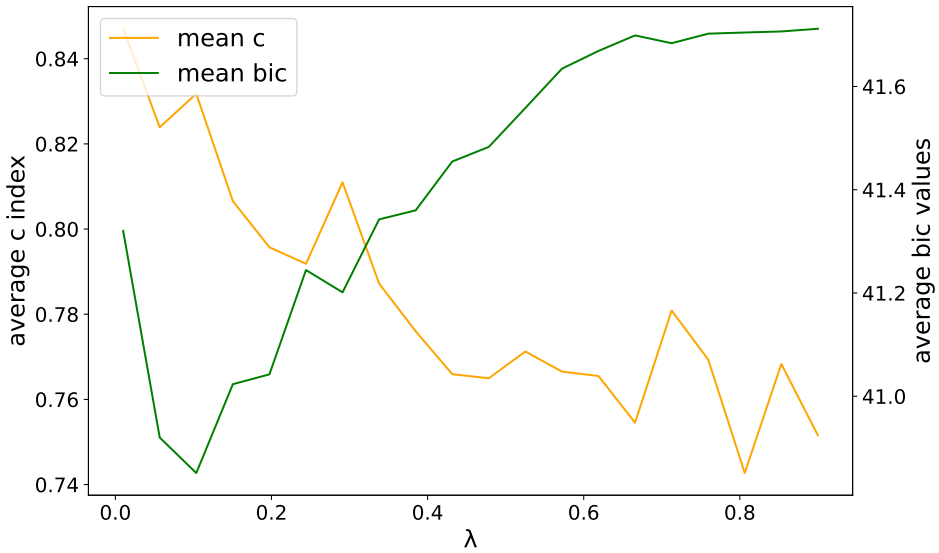



Figure 4.12: C-index and BIC values from 5-fold cross-validation using recurrent injuries from Team A

much linear impact the changes have on the covariates.

In terms of survival probability, we present how each covariate impacts the survival outcome over time using different covariate values. The

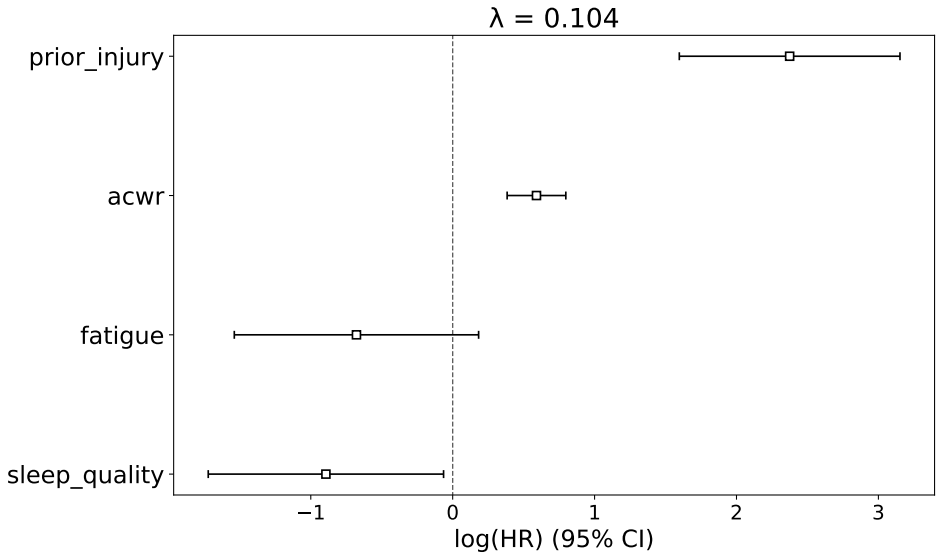



Figure 4.13: CPH analysis with optimal penalty term and covariates using recurrent injuries from Team A

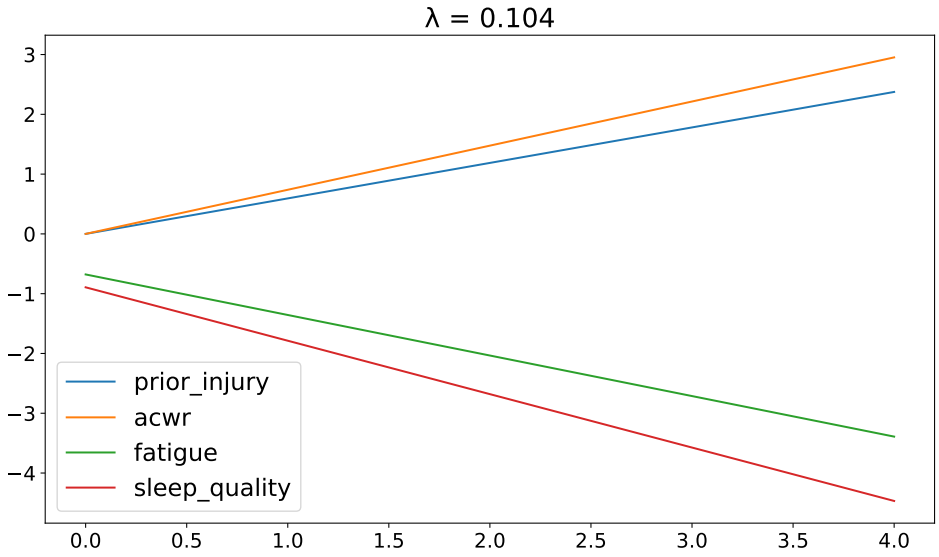
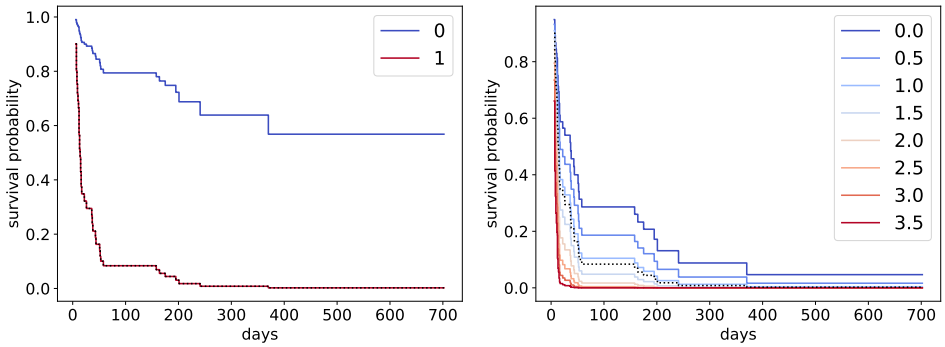



Figure 4.14: Injury risk functions for optimal covariates using recurrent injuries from Team A

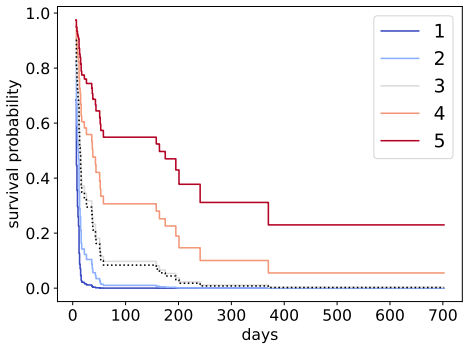
results are shown in Figure 4.15. The x-axis represents time in days, and the y-axis represents the survival probability, starting at 1.0. The plotted lines represent different values for the covariates. Prior injuries highly affect a player's probability of another injury, with the survival probability dropping to 10% by the first 50 days. A high ACWR of 3.5 drastically increases the chance of experiencing an injury, specifically by 100% in the first 10 days. A low fatigue score of 1, which indicates being very fatigued, decreases the survival probability by 100% in the first 50 days. We can also



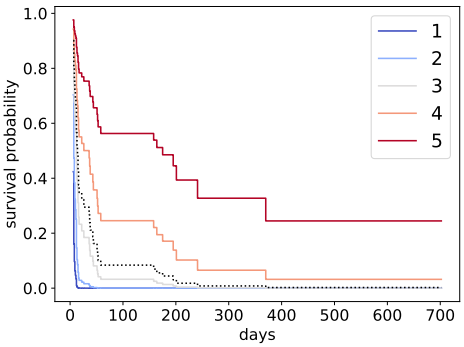

(a) Prior injury




(b) ACWR




(c) Fatigue




(d) Sleep quality

Figure 4.15: Partial effects on survival outcome of covariates using recurrent injuries from Team A

see a similar pattern for sleep quality.

Regarding our research questions, we can support some of our previous answers. For the question, "Does the size of the dataset affect the results?", we found in the previous experiments that this is true. Using smaller datasets, such as first injuries and Team B, we found that the estimated coefficients were unreliable and encountered issues with multicollinearity. In this experiment, we found that using first injuries results in very low variance, large confidence intervals or both, suggesting unreliable estimates. For instance, our results from using first injuries in this experiment suggest that increased sleep duration increases the risk of injury. From this experiment, we can support our previous conclusions that the size of the dataset does affect the results.

We also ask, "Does the number of covariates used in the multivariate models affect the results?". In Experiment 2, we found that using too many covariates led to multicollinearity issues for first injuries in Team B. In this experiment, we had to reduce the number of covariates as using all of them led to almost no variance in the coefficients and wide confidence intervals. By reducing the number of covariates, we got more variance in the coefficients of our covariates for first injuries. However, the confidence intervals were still quite large, and we assume the coefficients are unreliable. We support our previous answer that the number of

covariates affects the results, but keep in mind that reducing them might not help increase the confidence in the estimates, as it has more to do with the sample size being too small.

For the question regarding using first or recurrent injuries, we previously stated that using recurrent injuries is recommended for our case because including these represents the reality of soccer injuries, which often reoccur. We also found that including them provides more data to analyse, also enabling the use of more covariates. Recurrent injuries also introduce time-varying covariates in our analysis, as we can capture the changes from injury to injury over time. As discussed earlier, first injuries also result in uncertain estimates and suggestions we cannot rely on to be accurate. We came to the same conclusions in this experiment and decided to perform our regularisation with the CPH model using only recurrent injuries. Ultimately, this supports our previous answer that using recurrent injuries is recommended.

We still have two research questions that have not been answered, one of which is: "Should we use covariate values from the day of the injury or from its whole duration interval?". When comparing the results from Experiment 2 using day-of-the-event values with the results from this experiment using averaged values from whole intervals, we observe that averaged values result in higher variance in the coefficients. We also discuss the uncertainty in the day-of-the-event values, as we do not know if these values are reported before or after the injury, and injuries might affect factors such as mood, stress and sleep. Hence, we cannot determine if it is the covariates affecting injury or the injury affecting the covariates. Additionally, day-of-the-event values reflect more short-term effects on injury, while averaged values from whole intervals reflect more long-term effects. As we are more interested in the factors prior to injuries, the averaged values are a better fit for our research. With this in mind, we should use averaged values from whole duration intervals, as this provides more acceptable estimates and certainty and is more suitable for our research.

Our final question is, "What penalty term and threshold results in optimal feature selection?". As we have found in this experiment, the optimal penalty term is relatively low, precisely 0.104 in our results. This is the optimal value as it results in the CPH model with the lowest BIC and highest C-index. In other words, the model with the best performance. The threshold that results in optimal feature selection is 0.1, as it only allows for covariates with a certain amount of impact on injury risk.

In this experiment, we aimed to find the covariates in our dataset with the most significant impact on injury risk over time. We solved this by applying the CPH model to Team A's dataset using recurrent injuries with averaged covariate values from each duration interval and regularising with a penalty term of 0.104 and a threshold of 0.1. We found that the covariates with the highest impact on injury risk are prior injury, sleep quality, fatigue and ACWR.

## 4.6 Chapter Summary

This chapter presents our four experiments using the methods described in Chapter 3.

In Experiment 1, we validate our dataset for survival analysis using the univariate survival analysis models Kaplan-Meier, Weibull and Piecewise Exponential. We found that Kaplan-Meier and Piecewise Exponential had more accurate survival functions than Weibull, and their survival slopes matched our injury distribution.

In Experiment 2, we incorporate training load and wellness covariates from the dataset to estimate their significance on the injury outcome. We apply the CPH model using covariate values from the day of the event for both teams. We use first and recurrent injuries, with an added covariate, prior injury, for recurrent injuries. We include ACWR as this is a commonly used measure to assess injury risk. It also had low correlation with the other covariates. The wellness metrics chosen are fatigue, mood, readiness, sleep duration, sleep quality, soreness and stress. We apply the CPH model with the chosen covariates and find that the most informative covariates for first injuries are mood, fatigue and stress for Team A and stress and ACWR for Team B. For recurrent injuries, these are prior injury, ACWR and fatigue for Team A and prior injury, stress and ACWR for Team B. Overall, first injuries result in more variance in the coefficients than recurrent injuries but more uncertainty in their estimates due to a smaller sample size.

In Experiment 3, we treat our covariates as time-dependent, as they change over time. As in the previous example, we employ the CTV model for both teams. We find that using this type of model results in very low variance for the coefficients and conclude that this approach is not suitable for our analysis.

Lastly, in Experiment 4, we revisit the CPH model using averaged covariate values from the duration intervals. We do this for both first and recurrent injuries and find that first injuries either result in low variance in the coefficients or few possible covariates with large confidence intervals. Therefore, we proceed using only recurrent injuries, as this also has the benefit of representing changes in the covariates over time. We apply regularisation using L1 regression to extract the most significant covariates. We find the optimal penalty term for the regularisation by applying the CPH model with different values of $\lambda$, cross-validating each model and choosing the value yielding the lowest BIC and highest C-index, which is 0.104. The most significant covariates for Team A are prior injury, sleep quality, fatigue and ACWR, in that order. We also look at their risk functions and partial effects on outcome, showing that prior injuries drastically increase the chance of another injury. Decreased sleep quality and increased fatigue and ACWR highly increase the risk of injury.

In the next chapter, we discuss our overall insights and findings from the experiments and results in this chapter, as well as revisit our research questions, present potential use cases, and outline our limitations, next steps and contributions.

# Chapter 5

# Discussion

In the previous chapter, we presented the four experiments we conducted using the survival analysis techniques described in Chapter 3 and their respective outcomes.

In this chapter, we discuss the results of our experiments further and present the insights we derived from them. We revisit our research questions and answers and discuss the limitations of our study, its potential applications, and future work that can build on our research. Finally, we summarise our contributions to the field.

## 5.1 Insights

This section discusses the insights gained from our experiments using survival analysis to identify significant injury risk factors. We also discuss our findings regarding examining the impact of various factors, such as the use of first injuries versus recurrent injuries, the use of values from the day of the event versus the entire interval, and the use of time-dependent variables.

### 5.1.1 Prior Injury

We identified prior injury, sleep quality, fatigue and ACWR as the factors with the most effect on injury risk for Team A, in that order. Prior injury had the highest coefficient of approximately 2.4, indicating that having a prior injury increased the hazard by 240% compared to not having a prior injury. This means that athletes who had previously suffered an injury were more than twice as likely to sustain another injury than those who had not been injured before.

This finding is consistent with previous research, highlighting the influence of prior injuries on the likelihood of recurrent injuries [41]. It emphasises the importance of considering an athlete's injury history when assessing injury risk, as a history of injury may require more targeted injury prevention measures.

Additionally, this highlights the importance of incorporating recurrent events into survival analysis. Recurrent events like injuries are likely to

happen multiple times and may have different risk factors than the first occurrence. By accounting for recurrent events, the analysis can provide a more accurate assessment of injury risk and inform targeted prevention strategies.

### 5.1.2 Sleep Quality

Additionally, we found that sleep quality was another significant factor, with a coefficient of approximately -0.9, meaning a unit increase in sleep quality reduces the hazard by 90%. This coefficient was negative, implying that increased sleep quality decreases the likelihood of injury. This suggests that better sleep quality has a protective effect against injury and can be an essential factor in injury prevention strategies.

This finding is consistent with previous research showing the importance of adequate sleep in injury prevention and recovery [43]. Various factors can affect sleep quality, including lifestyle habits, stress, and workload. Therefore, promoting healthy sleep habits and addressing factors that can negatively impact sleep can be essential to injury prevention programs.

### 5.1.3 Fatigue

Fatigue appeared as another significant factor in injury risk, with a negative coefficient of approximately -0.7, indicating that an increase in the fatigue score reduced the hazard by 70%. Because fatigue is reported on a scale from 1 to 5, with 1 indicating "very tired" and 5 indicating "very fresh", increasing the score would mean reducing fatigue. Therefore, the results suggest that decreasing fatigue lowers the risk of injury.

Fatigue is important to injury risk because it can impair physical and cognitive function [30]. This can lead to a decreased ability to perform tasks, reduced reaction times, and decreased coordination, which can increase the risk of injury. Fatigue can also affect decision-making abilities, leading to poor judgment and increased risk-taking behaviours [56]. Additionally, fatigue can reduce the capacity for recovery and adaptation, making an individual more vulnerable to injury [47]. Therefore, understanding the impact of fatigue on injury risk is critical for developing effective injury prevention strategies.

### 5.1.4 ACWR

Lastly, the analysis showed that ACWR had a coefficient of approximately 0.6, indicating that increasing ACWR by one unit increased the hazard by 60%. As discussed in Section 2.6, ACWR is a commonly used measure of injury risk. Our findings support the concept that higher ACWR values significantly impact increased injury risk. However, when comparing the injury risk model proposed by Gabbett et al. in Figure 2.7 to our Figure 4.15 with ACWR's effect on the outcome, there is a difference in how lower ACWR values affect the injury outcome. Gabbett suggested that ACWR values lower than 0.8 increase the risk of injury, whereas our

findings showed that lower ACWR values decreased the risk of injury. Therefore, we cannot conclusively support using Gabbett's injury risk model to evaluate injury risk based on ACWR. This is consistent with the points expressed by Wang et al. [63], as we discussed in Section 2.6, and our results contribute to this scepticism regarding Gabbett's injury risk model.

Additionally, as discussed in 2.6, the ACWR injury risk model by Gabbett [25] only focuses on one single factor and provides only a snapshot of the factors affecting an athlete's health, performance and injury risk. As Bahr and Holm [5] state, focusing only on ACWR does not represent the complex nature of sports injuries. Therefore, it is essential to examine multiple injury risk factors to provide a more comprehensive understanding of an athlete's injury risk. As we found that prior injuries, sleep quality and fatigue have more impact on injury risk than ACWR, we prove how essential it is to investigate multiple factors and other factors than ACWR.

### 5.1.5 First Injuries vs Recurrent Injuries

For Experiments 2 and 4, using the CPH model, we found that the coefficients for first injuries had higher variance than those for recurrent injuries, indicating that covariates of first injuries have more impact on the injury risk. However, these coefficients had large confidence intervals, which suggests that the estimates of the coefficients are unreliable [7]. An explanation for this could be that the dataset for first injuries is smaller than for recurrent injuries and that there is not enough data to estimate these coefficients accurately. Additionally, we noticed a significant amount of data missing at the beginning of the dataset, indicating that there were few reports before the first injuries. This results in an even smaller dataset for first injuries, further affecting the results.

One possible explanation is that first injuries are more unpredictable and can be influenced by various factors. First injuries may be more likely to occur due to random chance or unexpected circumstances. In contrast, recurrent injuries may be influenced more by underlying physical and psychological factors that are more stable over time. Additionally, first injuries may be more mixed in severity, location, and causes, leading to more significant variability in the effects of different covariates. On the other hand, recurrent injuries may be more similar in nature and severity, which could result in more even effects of covariates for different cases. However, we cannot know if these first injuries are actually their first injury or simply the first injury reported during the observation period. Most likely, these soccer players have already experienced injuries prior to the study.

There was more variance and smaller confidence intervals for recurrent injuries, including all covariates, than for first injuries for both teams. Especially prior injury and sleep quality had a higher variance in the coefficients. This is likely due to the larger sample size of recurrent injuries than first injuries, allowing for a more stable estimation of the coefficients.

However, sleep quality for Team A and Team B had different effects.

For Team A, the coefficient suggested that increased sleep quality could reduce injury risk, but for Team B, the coefficient suggested otherwise. That improved sleep quality could increase injury risk. As discussed in Section 5.1.2, sleep quality is essential in injury prevention and recovery [43]. Therefore, we argue that the results using data from Team A are more accurate than those from Team B. Also, the confidence intervals for Team B using recurrent injuries are quite wide, so these coefficients have a high level of uncertainty.

As we incorporated recurrent injuries into our analysis, we discovered that many players on both teams had experienced multiple injuries. Given that soccer is one of the sports with the highest injury rates [54], this finding is not surprising. This also enabled us to introduce the prior injury covariate, which allowed us to explore the impact of prior injuries on the outcome. By including recurrent injuries in our dataset, we can accurately represent the reality of soccer and conduct more precise analyses, leading to more reliable results better suited for the team and the sport.

When analysing first injuries only, we may miss important information about how the covariates change over time and how they affect injury risk in the long term. By including recurrent injuries, we can track the changes in covariates before each injury and, therefore, better understand how changes in these factors can affect injury risk in the long term.

From our experiments, we have found that first injuries have higher variability in the coefficients compared to recurrent injuries but are more unreliable due to the small size of the dataset. Recurrent injuries provide more data and more reliable estimates. Additionally, including recurrent injuries depicts the reality of soccer injuries and how prior injury history may affect other injuries. Recurrent injuries also allow for time-varying analysis and can better capture the long-term effects of the covariates. Overall, studying recurrent injuries in addition to first injuries may provide a better understanding of the underlying risk factors.

### 5.1.6 Day-Of-The-Event Covariates vs Averaged Covariates

In Experiment 2, we used training load and wellness covariates from the day of the event, while in Experiment 4, we used the average covariate values from whole durations. When comparing the results for first injuries, we found that the averaged covariate values have larger confidence intervals than the day-of-the-event values. This indicates that the coefficients for the averaged values are unreliable. However, when using recurrent events, the averaged covariate values result in higher coefficient variance than the day-of-the-event values.

An explanation for this could be that day-of-the-event values capture more of the immediate effects on injury risk, which is more appropriate when using first injuries where there is less data prior to the event. On the other hand, averaged values over longer durations capture more of the long-term effects, which is more appropriate for recurrent events, as they capture effects over time. Therefore, it depends on whether the interest is to find the short-term or long-term effects. In our research, the combination

of recurrent injuries and averaged covariate values provide more benefits, as we have small datasets for first injuries.

A possible reason for the averaged covariate values resulting in such unreliable estimates for first injuries could be the loss of important information about changes in these values over time. For example, a player's wellness score might fluctuate during the duration of the interval, with high and low values occurring at different times. Averaging the values over the whole interval can result in a loss of this variability, leading to a less accurate representation of the player's true state during the interval. This loss of information can result in less reliable estimates of the coefficients, as the averaged values may not accurately reflect the true relationship between the covariates and injury risk. Therefore, averaged covariates from duration intervals should be combined with recurrent events as this represents the changes in the covariates over time.

There is a degree of uncertainty in the day-of-the-event values, as we do not know if these values are reported before or after the injury. Injuries might affect mood, stress and fatigue on the same day. Therefore, we cannot determine if the covariates affect injury or if the injury affects the covariates. Due to this uncertainty, using the averaged covariates may lead to more realistic estimates.

In summary, the choice of using either day-of-the-event or average covariate values depends on the nature of the covariate and the problem being investigated. For first injuries where short-term factors have more significance, such as mood, soreness and stress, day-of-the-event values may be more appropriate. Average covariate values over longer durations may be more appropriate for recurrent injuries where long-term factors are more significant, such as prior injuries, sleep quality and training load. One should keep in mind that averaging covariate values over longer durations could lead to a loss of information about the changes over time and should therefore be combined with recurrent events to represent these changes. Additionally, there is some uncertainty in using day-of-the-event values, as we do not know if they are reported prior to the injury or after.

### 5.1.7 Time-Dependent Covariates

As the data we use in this research is time series, we wanted to investigate the use of time-varying covariates. We applied the CTV model in Experiment 3. However, the model resulted in a very low variance between the coefficients, indicating that none of the covariates impacts injury risk over time. We observed that the confidence intervals of the coefficients are quite wide, indicating uncertainty in the estimated coefficients, likely due to the relatively small dataset size. If the sample size is small, the model may not be able to detect small but meaningful effects of the covariates and cannot detect significant associations between the covariates and injury risk.

Another possibility is that the time-varying covariates are not strongly associated with injury risk. However, in Experiment 4, we used a different approach to capture the time-varying covariates by combining recurrent

events with averaged covariate values. By doing this, we were able to capture the changes in the average from event to event in the same individual. This solution may not include the day-to-day changes as in Experiment 3 but includes changes in the intervals during the observation period. We found that the estimates of the coefficients were more certain and that the proposed significant covariates, prior injury, sleep quality, fatigue and ACWR made sense in terms of previous research on injury risk factors, as discussed in Sections 5.1.1, 5.1.2, 5.1.3 and 5.1.4. We cannot determine if these chosen covariates have long-term effects or if we truly capture their changes over time. However, we can propose this as an alternative to incorporating changes in the covariates over time.

An advantage to the solution in Experiment 4 is that it satisfies the proportional hazard assumption. As discussed in Section 3.2.4, the proportional hazard assumption is fundamental in survival analysis, stating that the ratio of hazards for any two groups being compared is constant over time. The assumption is important because violating it can lead to biased and unreliable results. If the effect of the covariate on the hazard change over time, the estimated hazard ratio may not accurately reflect the true relationship between the covariate and the hazard. Ensuring that the proportional hazards assumption is met provides more confidence in the results of the analysis. The CTV model does not meet the assumption because it allows for time-varying covariates, which means that a covariate's effect on an event's hazard can change over time. This could be a reason for the poor results in Experiment 3, and indicates that these results are unreliable. Therefore, using a solution such as in Experiment 4 can provide more confidence in the results than the time-varying model in Experiment 3.

Ultimately, the CTV model did not provide sufficient information about changes in the covariates over time, likely due to small sample sizes. However, a possibility to capture the changes in training load and wellness over time is to combine recurrent injuries with averaged covariate values from the durations. This solution provided more accurate results and satisfied the proportional hazard assumption, unlike the CTV model.

## 5.2 Revisiting the Research Questions

In this section, we revisit the research questions presented in Section 1.2 and present our answers based on our findings. We have seven smaller research questions that are extracted from our main research question. At the end of this section, we pull all strings together and use the answers from our smaller questions to answer our overall research question.

**RQ1. How should missing data entries be handled?** Missing data at the beginning and end of a player's dataset should be cut off. If there are injuries missing, this is solved through censoring. Missing covariate values scattered randomly throughout the dataset should be replaced with zeros if the multivariate model uses values from the day of the event, as removing

rows where these occur would remove valuable data. For multivariate models using averages, the missing values can be left as they are, as they are not considered in the mean calculations.

**RQ2. Can we use a univariate survival model to validate our dataset?** Univariate survival models can validate our dataset for survival analysis by comparing the injury distribution with the survival functions of models. The matching slopes of the survival functions and injury distribution indicated that our dataset is reliable and could be used for further analysis.

**RQ3. Does the size of the dataset affect the results?** The size of the dataset does impact the results significantly, and in the case of a small dataset, identifying the most significant covariates can be challenging. Results from smaller datasets, such as for Team B and first injuries, result in unreliable estimates of the coefficients.

**RQ4. Does the number of covariates used in the multivariate models affect the results?** Including too many covariates in a multivariate model with a smaller dataset can result in coefficients that are very close to zero, indicating that they may not have a meaningful impact on the outcome. Including too many covariates can also lead to issues with multicollinearity. This suggests that reducing the number of covariates can lead to more accurate results and a better understanding of the most significant factors contributing to the outcome of interest.

**RQ5. Should we use first injuries or recurrent injuries?** Including recurrent injuries is recommended when analysing injury risk in soccer players as it reflects the reality of the sport where multiple injuries are common. By considering recurrent injuries in our analysis, we had a larger dataset to work with, which led to more accurate results and a better understanding of how prior injuries may affect the likelihood of other injuries. Additionally, including recurrent injuries allows for capturing changes in the covariates over time.

**RQ6. Should we use covariate values from the day of the injury or from its whole duration interval?** Using averaged covariate values from whole durations is recommended if the aim is to investigate long-term effects on injury risk and the dataset includes recurrent injuries. Using day-of-the-event values is recommended if the aim is to investigate short-term or immediate effects on injury risk and the focus is on first injuries only. Our research prefers using averaged covariate values combined with recurrent injuries as it provides a larger dataset and leads to more accurate estimates.

**RQ7. What penalty term and threshold results in optimal feature selection?** The optimal penalty term based on BIC and C-index measurements

is 0.104. The optimal threshold is 0.1, which only allows for the most informative covariates.

We use the discussions from the respective sub-questions above to answer our overall research question:

> *How can we extract injury risk factors from training load and wellness data from elite female soccer teams using survival analysis?*

In this respect, extracting injury risk factors from an elite female soccer team using survival analysis can be done by applying a CPH model with regularisation using selected subjective training load and wellness covariates. It is recommended to include recurrent injuries, as multiple injuries are common in soccer, and this reflects the real nature of the event. Additionally, recurrent events also provide time-varying analysis. By using covariate values from whole duration intervals, we provide a more comprehensive picture of the factors leading up to an injury than day-of-the-event values. The optimal penalty term should be located by applying L1 regularisation with different $\lambda$ values and choosing the value resulting in the lowest BIC and highest C-index scores. The resulting optimal model extracts the most important risk factors for injury in the team.

## 5.3 Potential Use Cases

This section explores potential applications and uses cases for our solution, showcasing how it can be applied to different scenarios. By highlighting its versatility and adaptability, we aim to demonstrate our solution's wide range of possibilities to researchers, coaches, and other stakeholders interested in injury prevention.

**Application to Soccer Dashboard**: In Section 2.5, we introduced Soccer Dashboard, a web-based tool for data visualisation and analysis of pmSys and SoccerMon data. While the application provides valuable resources for players, coaches, team staff, and researchers, there is limited analysis and visualisation of injuries. Thus, a potential application of our solution is to incorporate it into Soccer Dashboard as a tool for analysing injury risk factors. The interactive nature of the dashboard allows users to explore various hyperparameters, covariates, teams and periods, making it a valuable addition to injury analysis, prevention and management.

**Development of injury prevention programs**: Our solution presents the opportunity to design tailored injury prevention programs for soccer teams. By identifying injury risk factors, teams and players can focus on reducing certain factors while increasing others to prevent injuries. For example, if high training loads are identified as a significant risk factor, coaches and trainers can adjust training programs to reduce the chance of injury. Similarly, if a lack of sleep is identified as a risk factor, teams and players can work on improving sleep habits to reduce injury risk.

**Application to other teams**: The analysis conducted in this research is based on data from two Norwegian elite soccer teams. However, the

methods used and the insights gained from the analysis can also be applied to other soccer teams. The approach taken in this research applies to any soccer team where data on wellness, training load, and injuries is collected. Our solution can identify injury risk factors for other teams in different leagues and countries.

**Application to other sports**: In our research, we use data from soccer teams to identify injury risk factors. However, the data we use can be collected from any sport. Injuries are common in all sports, making our solution potentially useful for injury prevention in other sports. Moreover, our approach of analysing teams as a whole, rather than focusing solely on individual players, could be useful for team sports. In team sports, an individual player's performance could be dependent on the team's performance as a whole. Therefore, analysing the team as a whole could provide valuable insights into the factors contributing to the team's injury risk as a unit.

## 5.4 Limitations

In this section, we discuss some of the limitations of our experiments and solutions, such as the dataset size and missing data, how we differentiate injuries and perform time-dependent analysis, and the few multivariate models we use for analysis.

### 5.4.1 Dataset Size

While the Soccermon dataset is a rich data source for injury analysis, the limited number of injuries reported during the period of interest, 2020 to 2021, could be a limitation of our analysis. External factors, such as COVID-19 restrictions on training and matches, could have impacted the number of reported injuries during this period. As a result, our analysis might differ from the teams' typical injury patterns.

### 5.4.2 Missing Data

As discussed in Section 4.1.2, missing data is a common challenge in real-world datasets, and our dataset for this thesis is no exception due to subjective reporting. This limitation presents several issues for our analysis.

To address missing data, we trim off the beginning and end of player datasets, where larger amounts of data are missing, perhaps due to late reporting or players switching teams. We use ACWR to measure valid entries, remove data up to the first valid entry, and cut off from the last valid entry to the end. However, this approach may inadvertently remove valuable injury data that could be useful in our analysis. For instance, Team A had reported 13 first injuries, but when cutting off missing covariate values, we were only left with 6 injuries.

Moreover, our dataset contains random missing data points. If their corresponding rows were to be removed, we would have a small dataset

that could negatively affect our results. In some cases, these rows with missing data points may still contain valuable information, such as instances that report an injury but are missing wellness values. To address this issue, we fill random missing data points with zeros. However, this could skew our results by creating artificially low training load and wellness metrics values.

### 5.4.3 Differentiating Injuries

In our dataset, players may have reported the same injury over multiple days, and there is no differentiation between new and old injuries. To avoid including duplicate injuries, we set a 5-day limit between injuries. However, this could lead to removing new injuries within that window.

Additionally, we could not distinguish between impact and overuse injuries, which may limit our ability to identify the specific types of injuries being examined. Impact injuries can be more random and unpredictable, while overuse injuries can often be more predictable.

### 5.4.4 Time-Dependent Analysis

The dataset used in this thesis is based on time-series data, where the variables under analysis change over time. Given the time-dependency of the data, it is essential to account for these changes when analysing the data. In Experiment 3, we aimed to address this by using the CTV model that accounts for time-dependent variables. However, the results were inconclusive. When using the time-varying model, we only applied it to first injuries, a small dataset that may have been the reason for the inadequate results. By not including recurrent injuries, we may have missed the potential of the model, and it could have been a useful tool for time-varying analysis if we had included this data.

We explored an alternative approach by using recurrent events and averaging the values within each event's duration interval. While this approach captures the changes in the variables over time, it may have a potential limitation. Specifically, there may need to be more recurrent events for each player, and averaging the values within each interval may reduce the changes over time to a single value, potentially oversimplifying the analysis.

### 5.4.5 Few Multivariate Models

In our experiments using multivariate models, we only use the CPH model and its extension, the CTV model. While these are widely used and well-established methods in survival analysis, other survival analysis models could potentially provide different insights or be a better fit for the data. By limiting the analysis to these two models, it is possible that some essential aspects of the data are not captured, or the results are not as accurate as they could be.

## 5.5 Future Work

While our solution provides insights into injury risk factors in soccer teams, several areas remain for potential future work and improvements. In this section, we outline some of the next steps that could be taken to expand and enhance our solution.

**Use data from 2022**: As described in Section 5.4.1, one of our limitations is that we have a small dataset with few injuries. Therefore, a next step would be to include data from 2022, providing a larger dataset with more injuries to analyse. This could improve the reliability of the analysis and possibly provide better results for the CTV model. Additionally, the dataset from 2022 contains more information about the injuries, such as if the injury is from impact or overuse and if the injury is new or a previous injury.

**Use last values for injuries missing covariates**: One of our limitations is that we cut off the beginning and end of the players' datasets with missing covariate values, as described in Section 5.4.2, resulting in losing valuable injury data. An improvement could be replacing these missing values with the first or last present values. For instance, instead of removing a player's injury occurring at the end of the dataset, we would replace the missing training load and wellness values with the previous covariate values for that player.

**Use interpolated values for missing data points**: As presented in Section 5.4.2, another limitation of our solution is that we fill randomly scattered missing data points with zeros. In a next step, we would explore interpolation, meaning we would fill missing data points with values based on calculations from neighbouring data points. This could potentially avoid the issue of skewing our results by creating artificially low values for training load and wellness metrics.

**Differentiate injuries by location**: In our solution, we distinguish between separate injuries using a 5-day limit between each, as discussed in Section 5.4.3. This could result in losing separate injuries that occur within the same window. A possibility is to use the injury location to differentiate them instead. However, an injury could occur in the same location and be a different injury, but this is less likely than it is to skip injuries within the same 5-day window.

For future work, including data from 2022 would provide this type of information, meaning injuries are separable by type and occurrence. Using this data could lead to a more accurate injury risk analysis.

**Use recurrent injuries in time-varying model**: As mentioned in Section 5.4.4, we have not explored the full potential of the CTV model, as we only used first injuries, which is a small dataset. One of our next steps would be incorporating recurrent injuries with time-varying analysis to obtain more informative and accurate results.

**Explore other models**: As we only employed the CPH model and the CTV model for multivariate analysis, exploring other survival analysis models in future research could be worthwhile to ensure a more comprehensive analysis of the data. Several other survival analysis models can be explored, such as the Accelerated Failure Time (AFT) model, the Weibull

AFT model and frailty models, among others [36]. Exploring these alternative models could provide additional insights and improve the accuracy of the analyses.

**Use the number of previous injuries**: In our solution, we introduce the prior injury covariate as a binary variable indicating whether a player has had an injury previously in the observation period. However, counting the number of previous injuries could result in a more accurate estimate of the prior injury variable. This approach would provide a better understanding of how the degree of the number of injuries influences the injury risk and could lead to more targeted injury prevention strategies. Therefore, exploring the use of a count variable for prior injuries would be one of our next steps.

**Use other factors**: Including other factors, such as psychological and nutritional factors, can provide a more comprehensive understanding of injury risk factors. Psychological factors can affect an athlete's physical and emotional well-being, which could influence injury risk. Nutritional factors like diet and hydration can also impact an athlete's physical performance and recovery. Therefore, incorporating these factors into our solution could lead to a more accurate assessment of injury risk and a more effective injury prevention strategy.

Furthermore, it may be valuable to include illnesses as events, as they can also significantly impact athlete performance and availability. By including data on illnesses, researchers could identify factors contributing to the increased risk of illness and develop interventions to reduce that risk.

**Focus on specific injury events**: Besides analysing injury events in general, future work could focus on specific types of injuries, such as knee injuries, head injuries, or other common injuries in the sport being studied. By analysing injury events at a more granular level, researchers could identify risk factors and preventative measures specific to certain types of injuries rather than just injuries in general.

## 5.6 Contributions

Our contributions from this thesis include open-source software and insights for both computer and sports science. All the software used in this research is accessible under this GitHub repository: https://github.com/simula/pmsys. We contribute to the field of computer science by providing alternative methods to evaluate risk and identify significant factors in a dataset. Additionally, we provide techniques for data processing, handling missing data and small datasets, survival analysis, feature selection and regularisation. We also extend the application of survival analysis beyond its traditional use in medical research to a sports science context, a relatively new and growing area of research.

By analysing training load and wellness data using survival analysis, we have identified the factors that contribute to injury risk in a soccer team. This can benefit players, coaches, and medical staff by better understanding the factors that affect injury risk and enabling them to develop effective

injury prevention strategies. Our method applies not only to soccer but also to other sports where injury risk is a concern, making it a versatile tool for injury prevention.

Furthermore, our research contributes to sports science by identifying injury risk factors specific to female elite soccer teams. While previous research has focused on injury risk factors in male soccer teams, there is limited research on female soccer teams. Our findings aid in filling this knowledge gap and improving injury prevention strategies for female soccer teams.

## 5.7 Chapter Summary

In this chapter, we present our key insights from our experiments using survival analysis techniques and discuss the significance of our findings in the context of our original research questions. We present several potential applications demonstrating the significance of our research, such as using our solutions in the interactive Soccer Dashboard, in the development of injury prevention programs and applying it to other teams and sports. We elaborate on the limitations of our work, such as dataset size, missing data, differentiating injuries, time-varying analysis and the use of few models. Additionally, we outline potential future work, such as using newer datasets, handling missing data differently, improving time-varying analysis and exploring other models and potential risk factors. We conclude with an overview of our contributions to the fields of computer science and sports science.

# Chapter 6

# Conclusion

In this thesis, we experiment with survival analysis techniques to identify significant injury risk factors in subjective training load and wellness data from elite female soccer teams. We investigate the effect of dataset size and the number of covariates used, and how missing data should be handled. The univariate survival analysis models Kaplan-Meier, Weibull and Piecewise Exponential are used to inspect if the dataset is suited for survival analysis. We use the Cox Proportional Hazards Model to estimate the significance of each covariate, and we apply regularisation with different penalty terms to extract the most essential factors. We also investigate the use of the Cox Time-Varying Model with time-dependent covariates. Additionally, we compare the use of first injuries versus recurrent injuries and the use of day-of-the-event values versus using averaged values from all days prior.

Our results showed that using the Cox Proportional Hazards Model with regularisation using a low penalty term and a threshold of 10% resulted in identifying the most critical injury risk factors: prior injury, sleep quality, fatigue and ACWR. We found that using smaller datasets impacts the results as they reduce the possible covariates, and the estimates for the covariates are unreliable. More extensive datasets, however, result in more certain estimates and injury risk factors that make sense based on previous research while also allowing for more covariates. When using multivariate models, the number of covariates used does affect the outcome, as using too many can result in issues with multicollinearity. Missing data at the start and end of a player's dataset is handled by cutting them off. Randomly scattered missing data points are replaced by zeros, except for the case of using averaged values from whole durations, where the missing data points are left as is. We found that using recurrent injuries reflects more of the real world, as it is common for soccer players to encounter multiple injuries. It also has the benefit of providing more data for analysis. Averaged covariate values from whole duration intervals are preferred as they can be combined with recurrent injuries and capture changes in the values over time.

In terms of computer science, this thesis contributes to the concepts of data processing, handling missing data and small datasets, survival

analysis, feature selection and regularisation. Our research also provides alternative methods to evaluate risk and identify significant factors in a dataset. We also extend the application of survival analysis beyond its traditional use in medical research to a sports science context, a relatively new and growing area of research.

This research contributes to the field of sports science by identifying risk factors associated with injury in soccer teams, which can benefit players, coaches and medical staff by providing a better understanding of the factors and enabling them to develop injury prevention strategies. Additionally, we contribute to filling the knowledge gap and the need for more research in women's soccer.

Looking ahead, our findings and methods can be used in any sport involving injury prevention. Additionally, our research can be extended by exploring other factors affecting injury risk, specifying the type of injury of interest, and considering other datasets, leagues, and countries. Ultimately, our research aims to contribute to developing injury prevention programs in soccer and other sports, benefiting players, coaches, and medical staff.

# Bibliography


[1] R. Aicale, D. Tarantino and N. Maffulli. 'Overuse injuries in sport: a comprehensive overview'. In: *J Orthop Surg Res* 13.1 (2018), p. 309. DOI: 10.1186/s13018-018-1017-5.

[2] Aylin Alin. 'Multicollinearity'. In: *WIREs Computational Statistics* 2.3 (2010), pp. 370–374. DOI: https://doi.org/10.1002/wics.84. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wics.84. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.84.

[3] Leila DAF Amorim and Jianwen Cai. 'Modelling recurrent events: a tutorial for analysis in epidemiology'. In: *International Journal of Epidemiology* 44.1 (Dec. 2014), pp. 324–333. ISSN: 0300-5771. DOI: 10.1093/ije/dyu222. eprint: https://academic.oup.com/ije/article-pdf/44/1/324/14152617/dyu222.pdf. URL: https://doi.org/10.1093/ije/dyu222.

[4] Renato Andrade et al. 'Is the Acute: Chronic Workload Ratio (ACWR) Associated with Risk of Time-Loss Injury in Professional Team Sports? A Systematic Review of Methodology, Variables and Injury Risk in Practical Situations'. In: *Sports Medicine* (2020), pp. 1613–1635. DOI: 10.1007/s40279-020-01308-6.

[5] R Bahr and I Holme. 'Risk factors for sports injuries — a methodological approach'. In: *British Journal of Sports Medicine* 37.5 (2003), pp. 384–392. ISSN: 0306-3674. DOI: 10.1136/bjsm.37.5.384. eprint: https://bjsm.bmj.com/content/37/5/384.full.pdf. URL: https://bjsm.bmj.com/content/37/5/384.

[6] Daniel Berrar. 'Cross-Validation'. In: *Encyclopedia of Bioinformatics and Computational Biology*. Ed. by Shoba Ranganathan et al. Oxford: Academic Press, 2019, pp. 542–545. ISBN: 978-0-12-811432-2. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20349-X. URL: https://www.sciencedirect.com/science/article/pii/B978012809633820349X.

[7] Rebecca Bevans. 'Understanding Confidence Intervals | Easy Examples & Formulas'. In: *Scribbr* (Nov. 2022). URL: https://www.scribbr.com/statistics/confidence-interval/.

[8] T. Bibson. 'Sports injuries'. In: *Baillière's Clinical Rheumatology* 1.3 (1987). Epidemiological, Sociological and Environmental Aspects of Rheumatology, pp. 583–600. ISSN: 0950-3579. DOI: https://doi.org/10.1016/S0950-3579(87)80046-8. URL: https://www.sciencedirect.com/science/article/pii/S0950357987800468.

[9] Chris Bodenner. *Why Aren't Women's Sports as Big as Men's? Your Thoughts*. 2015. URL: https://www.theatlantic.com/entertainment/archive/2015/06/women-and-sports-world-cup-soccer/395231/.

[10] Matthias Boeker and Cise Midoglu. 'Soccer Athlete Data Visualization and Analysis with an Interactive Dashboard'. In: *MultiMedia Modeling*. Ed. by Duc-Tien Dang-Nguyen et al. Cham: Springer International Publishing, 2023, pp. 565–576. ISBN: 978-3-031-27077-2. DOI: 10.1007/978-3-031-27077-2_44. URL: https://soccer-dashboard.simula.no/.

[11] Adam R Brentnall and Jack Cuzick. 'Use of the concordance index for predictors of censored survival data'. In: *Statistical methods in medical research* 27.8 (2018), pp. 2359–2373.

[12] Jason Brownlee. 'How to Configure k-Fold Cross-Validation'. In: *Machine Learning Mastery* (July 2020). URL: https://machinelearningmastery.com/how-to-configure-k-fold-cross-validation/.

[13] T G Clark et al. 'Survival analysis part I: basic concepts and first analyses'. In: *British Journal of Cancer* (2003). DOI: 10.1038/sj.bjc.6601118.

[14] D. R. Cox. 'Regression Models and Life-Tables'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202. DOI: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00899.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x.

[15] James P Curley and Oliver Roeder. 'English soccer's mysterious worldwide popularity'. In: *Contexts* 15.1 (2016), pp. 78–81.

[16] Cameron Davidson-Pilon. 'lifelines: survival analysis in Python'. In: *Journal of Open Source Software* 4.40 (2019), p. 1317. DOI: 10.21105/joss.01317. URL: https://doi.org/10.21105/joss.01317.

[17] P.J. Denning et al. 'Computing as a discipline'. In: *Computer* 22.2 (1989), pp. 63–70. DOI: 10.1109/2.19833.

[18] Tanujit Dey et al. 'Survival analysis—time-to-event data and censoring'. en. In: *Nature Methods* 19.8 (Aug. 2022), pp. 906–908. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01563-7. URL: https://doi.org/10.1038/s41592-022-01563-7.

[19] *FFRC: UIT*. URL: https://uit.no/research/ffrc.

[20] Caroline Finch. 'A new framework for research leading to sports injury prevention'. In: *Journal of Science and Medicine in Sport* 9.1 (2006), pp. 3–9. ISSN: 1440-2440. DOI: https://doi.org/10.1016/j.jsams.2006.02.009. URL: https://www.sciencedirect.com/science/article/pii/S1440244006000235.

[21] *Forzasys: PmSys*. URL: https://forzasys.com/pmSys.html.

[22] Carl Foster et al. 'A new approach to monitoring exercise training'. In: *Journal of strength and conditioning research* 15.1 (Feb. 2001), pp. 109–115.

[23] Norges Fotballforbund. *Toppserien 2020 - Terminliste*. https://www.fotball.no/fotballdata/turnering/terminliste/?fiksId=169786.

[24] Norges Fotballforbund. *Toppserien 2021 - Terminliste*. https://www.fotball.no/fotballdata/turnering/terminliste/?fiksId=174541.

[25] Tim J Gabbett. 'The training—injury prevention paradox: should athletes be training smarter and harder?' In: *British Journal of Sports Medicine* 50.5 (2016), pp. 273–280. ISSN: 0306-3674. DOI: 10.1136/bjsports-2015-095788. eprint: https://bjsm.bmj.com/content/50/5/273.full.pdf. URL: https://bjsm.bmj.com/content/50/5/273.

[26] Tim J Gabbett et al. 'The athlete monitoring cycle: a practical guide to interpreting and applying training monitoring data'. In: *British Journal of Sports Medicine* 51.20 (2017), pp. 1451–1452. ISSN: 0306-3674. DOI: 10.1136/bjsports-2016-097298. eprint: https://bjsm.bmj.com/content/51/20/1451.full.pdf. URL: https://bjsm.bmj.com/content/51/20/1451.

[27] Danica N. Giugliano and Jennifer L. Solomon. 'ACL Tears in Female Athletes'. In: *Physical Medicine and Rehabilitation Clinics of North America* 18.3 (2007). Gender Specific Medicine: The Physiatrist and Women's Health, pp. 417–438. ISSN: 1047-9651. DOI: https://doi.org/10.1016/j.pmr.2007.05.002. URL: https://www.sciencedirect.com/science/article/pii/S1047965107000459.

[28] Stephanie Glen. *Univariate Analysis: Definition, Examples*. https://www.statisticshowto.com/univariate/. Accessed on April 19, 2023. n.d.

[29] Manu K Goel, Pardeep Khanna and Jugal Kishore. 'Understanding survival analysis: Kaplan-Meier estimate'. In: *International Journal of Ayurveda Research* 1.4 (2010), pp. 274–278. DOI: 10.4103/0974-7788.76794.

[30] Glenn Gunzelmann and Kevin Gluck. 'An Integrative Approach to Understanding and Predicting the Consequences of Fatigue on Cognitive Performance'. In: *Air Force Research Lab Mesa AZ Human Effectiveness Directorate* (Jan. 2009). Accession Number: ADA514141.

[31] Martin Hägglund et al. 'Injuries affect team performance negatively in professional football'. In: *British Journal of Sports Medicine* 47.12 (2013), pp. 738–742. DOI: 10.1136/bjsports-2013-092215.

[32] Charles R. Harris et al. 'Array programming with NumPy'. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[33] Will G Hopkins et al. 'Risk factors and risk statistics for sports injuries'. In: *Clinical Journal of Sport Medicine* 17.3 (May 2007), pp. 208–210. DOI: 10.1097/JSM.0b013e3180592a68.

[34] J. D. Hunter. 'Matplotlib: A 2D graphics environment'. In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[35] Håvard D. Johansen et al. 'Scalable Infrastructure for Efficient Real-Time Sports Analytics'. In: *Companion Publication of the 2020 International Conference on Multimodal Interaction*. ICMI '20 Companion. Virtual Event, Netherlands: Association for Computing Machinery, 2021, pp. 230–234. ISBN: 9781450380027. DOI: 10.1145/3395035.3425300. URL: https://doi.org/10.1145/3395035.3425300.

[36] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text. A Self-Learning Text, Third Edition*. 3rd ed. Statistics for Biology and Health. NY: Springer New York, 2011, pp. XV, 700. DOI: 10.1007/978-1-4419-6646-9.

[37] Ilari Kuitunen et al. 'Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review'. In: *BMC Musculoskeletal Disorders* 22.1 (2021), p. 489. ISSN: 1471-2474. DOI: 10.1186/s12891-021-04379-2. URL: https://doi.org/10.1186/s12891-021-04379-2.

[38] Rekha M. *MLmuse: Correlation and Collinearity — How they can make or break a model*. July 2019. URL: https://blog.clairvoyantsoft.com/correlation-and-collinearity-how-they-can-make-or-break-a-model-9135fbe6936a.

[39] Laurent Malisoux et al. 'Monitoring of sport participation and injury risk in young athletes'. In: *Journal of Science and Medicine in Sport* 16.6 (2013), pp. 504–508. ISSN: 1440-2440. DOI: https://doi.org/10.1016/j.jsams.2013.01.008. URL: https://www.sciencedirect.com/science/article/pii/S1440244013000285.

[40] Sam McCormick. 'Survival Analysis Part 1: The Weibull Model'. In: *Medium* (2018). URL: https://medium.com/utility-machine-learning/survival-analysis-part-1-the-weibull-model-5c2552c4356f.

[41] Willem H Meeuwisse et al. 'A dynamic model of etiology in sport injury: The recursive nature of risk and causation'. In: *Clinical Journal of Sport Medicine* 17.3 (May 2007), pp. 215–219. DOI: 10.1097/JSM.0b013e3180592a48.

[42] Cise Midoglu et al. *SoccerMon, A Large-Scale Multivariate Soccer Athlete Health, Performance, and Position Monitoring Dataset*. Open Science Framework (OSF). 2023. URL: https://doi.org/10.17605/OSF.IO/URYZ9.

[43] Matthew D. Milewski et al. 'Chronic Lack of Sleep is Associated With Increased Sports Injuries in Adolescent Athletes'. In: *Journal of Pediatric Orthopaedics* 34.2 (Mar. 2014), pp. 129–133. DOI: 10.1097/BPO.0000000000000151.

[44] K. Jarrod Millman and Michael Aivazis. 'Python for Scientists and Engineers'. In: *Computing in Science & Engineering* 13.2 (2011), pp. 9–12. DOI: 10.1109/MCSE.2011.36.

[45] Ahmed Naglah et al. 'Athlete-Customized Injury Prediction using Training Load Statistical Records and Machine Learning'. In: *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. 2018, pp. 459–464. DOI: 10.1109/ISSPIT.2018.8642739.

[46] Andrew A. Neath and Joseph E. Cavanaugh. 'The Bayesian information criterion: background, derivation, and applications'. In: *WIREs Computational Statistics* 4.2 (2012), pp. 199–203. DOI: https://doi.org/10.1002/wics.199. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wics.199. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.199.

[47] Mathieu Nédélec et al. 'Recovery in Soccer'. eng. In: *Sports Medicine* 42.12 (Dec. 2012), pp. 997–1015. ISSN: 1179-2035. DOI: 10.1007/BF03262308. URL: https://doi.org/10.1007/BF03262308.

[48] Rasmus Oestergaard Nielsen et al. 'Time-to-event analysis for sports injury research part 2: time-varying outcomes.' In: *British Journal of Sports Medicine* 53.1 (2019), pp. 70–78. DOI: 10.1136/bjsports-2018-100000.

[49] A Olayinka, Alfred Abiodun and Aliyu Ishaq. 'The Use of Cox and Piecewise Exponential Models in the Determination of Renal Failure'. In: 3 (Sept. 2020), pp. 91–100.

[50] Michał Oleszak. 'Feature Selection Methods and How to Choose Them'. In: (Apr. 2023). URL: https://neptune.ai/blog/feature-selection-methods.

[51] Sebastian Pölsterl. 'scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn'. In: *Journal of Machine Learning Research* 21.212 (2020), pp. 1–6. URL: http://jmlr.org/papers/v21/20-729.html.

[52] Jean-Baptist du Prel et al. 'Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications.' eng. In: *Deutsches Arzteblatt international* 106 (19 May 2009), pp. 335–9. ISSN: 1866-0452. DOI: 10.3238/arztebl.2009.0335.

[53] Jonas Ranstam and Jonathan A Cook. 'LASSO regression'. In: *British Journal of Surgery* 105.10 (Sept. 2018), p. 1348. DOI: 10.1002/bjs.10895.

[54] Pinyao Rui, Jill J Ashman and Akintunde Akinseye. 'Emergency Department Visits for Injuries Sustained During Sports and Recreational Activities by Patients Aged 5-24 Years, 2010-2016'. In: *National health statistics reports* 133 (2019), pp. 1–15.

[55] Anna Saw, Luana Main and Paul Gastin. 'Monitoring the athlete training response: Subjective self-reported measures trump commonly used objective measures: A systematic review'. In: *British journal of sports medicine* 50 (Oct. 2015). DOI: 10.1136/bjsports-2015-094758.

[56] Mitchell R. Smith et al. 'Mental fatigue impairs soccer-specific decision-making skill'. In: *Journal of Sports Sciences* 34.14 (2016). PMID: 26949830, pp. 1297–1304. DOI: 10.1080/02640414.2016.1156241. eprint: https://doi.org/10.1080/02640414.2016.1156241. URL: https://doi.org/10.1080/02640414.2016.1156241.

[57] Torbjørn Soligard et al. 'How much is too much? (Part 1) International Olympic Committee consensus statement on load in sport and risk of injury'. In: *British Journal of Sports Medicine* 50.17 (2016), pp. 1030–1041. ISSN: 0306-3674. DOI: 10.1136/bjsports-2016-096581. eprint: https://bjsm.bmj.com/content/50/17/1030.full.pdf. URL: https://bjsm.bmj.com/content/50/17/1030.

[58] Kathrin Steffen et al. 'ECSS Position Statement 2009: Prevention of acute sports injuries'. In: *European Journal of Sport Science* 10.4 (2010), pp. 223–236. DOI: 10.1080/17461390903585173. eprint: https://doi.org/10.1080/17461390903585173. URL: https://doi.org/10.1080/17461390903585173.

[59] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[60] Hans Van Eetvelde et al. 'Machine learning methods in sport injury prediction and prevention: a systematic review'. In: *Journal of Experimental Orthopaedics* 8.1 (2021), p. 27. ISSN: 2197-1153. DOI: 10.1186/s40634-021-00346-x. URL: https://doi.org/10.1186/s40634-021-00346-x.

[61] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[62] E. Verhagen. 'The cost of sports injuries'. In: *Journal of Science and Medicine in Sport* 13 (2010). 2010 Asics Conference of Science and Medicine in Sport "Hot topics in the tropics", 3-6 November 2010, e40. ISSN: 1440-2440. DOI: https://doi.org/10.1016/j.jsams.2010.10.546. URL: https://www.sciencedirect.com/science/article/pii/S1440244010007474.

[63] Chinchin Wang et al. 'Analyzing Activity and Injury: Lessons Learned from the Acute:Chronic Workload Ratio'. In: *Sports Medicine* 50 (2020), pp. 1243–1254. ISSN: 1179-2035. URL: https://doi.org/10.1007/s40279-020-01280-1.

[64] Chinchin Wang et al. 'Predicting Injury Risk Over Changes in Physical Activity in Children Using the Acute:Chronic Workload Ratio'. In: *American Journal of Epidemiology* 191.4 (Nov. 2021), pp. 665–673. ISSN: 0002-9262. DOI: 10.1093/aje/kwab280. eprint: https://academic.oup.com/aje/article-pdf/191/4/665/43019953/kwab280.pdf. URL: https://doi.org/10.1093/aje/kwab280.

[65] Michael L. Waskom. 'seaborn: statistical data visualization'. In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[66] Andrew Watson et al. 'Subjective well-being and training load predict in-season injury and illness risk in female youth soccer players.' eng. In: *British journal of sports medicine* 51.3 (Feb. 2017). PMID: 27919919, pp. 194–199. ISSN: 0306-3674. DOI: 10.1136/bjsports-2016-096584. URL: https://www.ncbi.nlm.nih.gov/pubmed/27919919.

[67] Theodor Wiik et al. 'Predicting Peek Readiness-to-Train of Soccer Players Using Long Short-Term Memory Recurrent Neural Networks'. In: *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. 2019, pp. 1–6. DOI: 10.1109/CBMI.2019.8877406.

[68] Xue Ying. 'An Overview of Overfitting and its Solutions'. In: *Journal of Physics: Conference Series* 1168.2 (Feb. 2019), p. 022022. DOI: 10.1088/1742-6596/1168/2/022022. URL: https://dx.doi.org/10.1088/1742-6596/1168/2/022022.

[69] Zhigang Zhang. 'Semi-parametric regression model for survival data: graphical visualization with R'. In: *Annals of translational medicine* 4.23 (2016), p. 461. DOI: 10.21037/atm.2016.08.61.

[70] Zhongheng Zhang et al. 'Time-varying covariates and coefficients in Cox regression models'. In: *Annals of Translational Medicine* 6 (Apr. 2018), pp. 121–121. DOI: 10.21037/atm.2018.02.12.

# Appendix A

# Supplementary Figures For Team B

## A.1 Experiment 4 - Cox Proportional Hazards Model With Regularisation

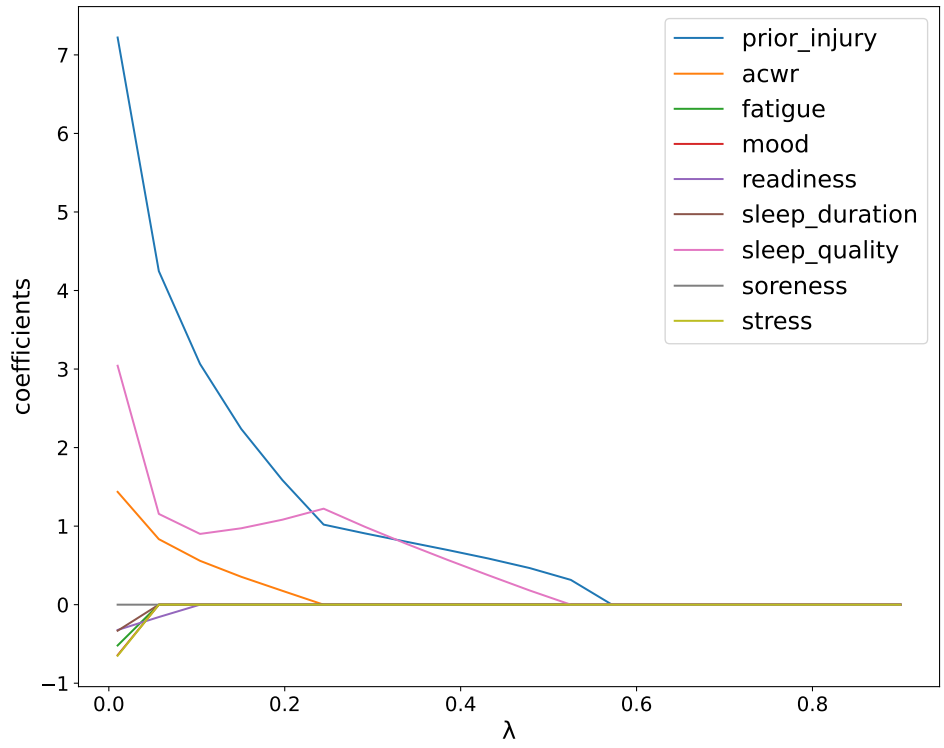



Figure A.1: CPH analysis with L1 regularisation using recurrent injuries from Team B

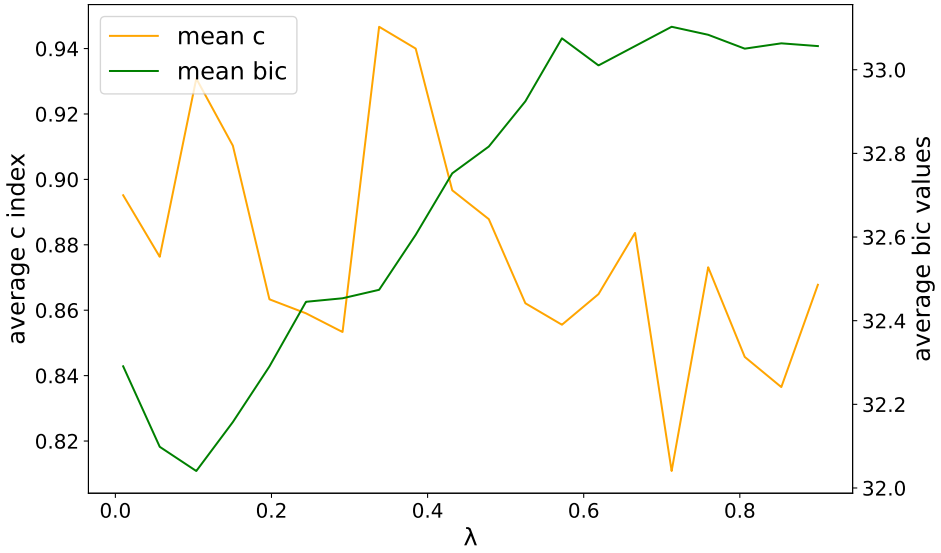



Figure A.2: C-index and BIC values from 5-fold cross-validation using recurrent injuries from Team B

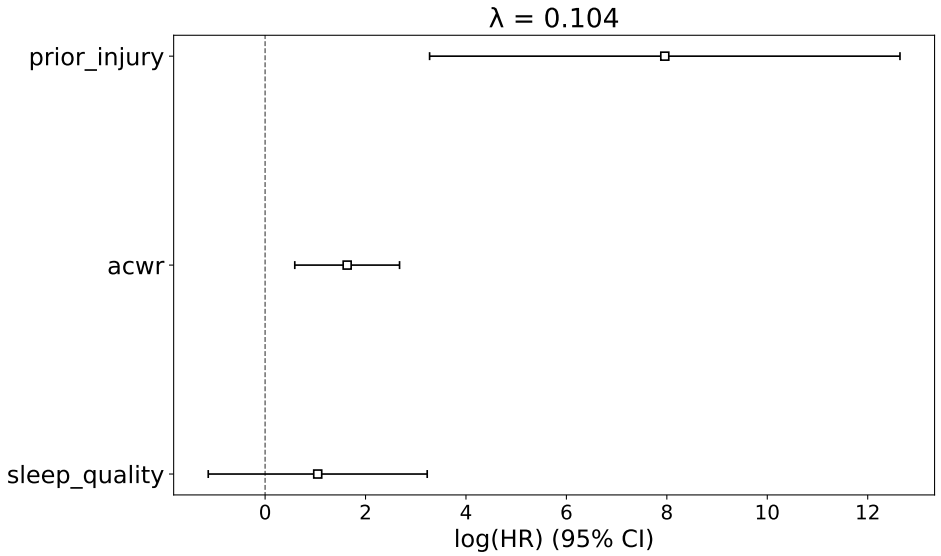



Figure A.3: CPH analysis with optimal penalty term and covariates using recurrent injuries from Team B

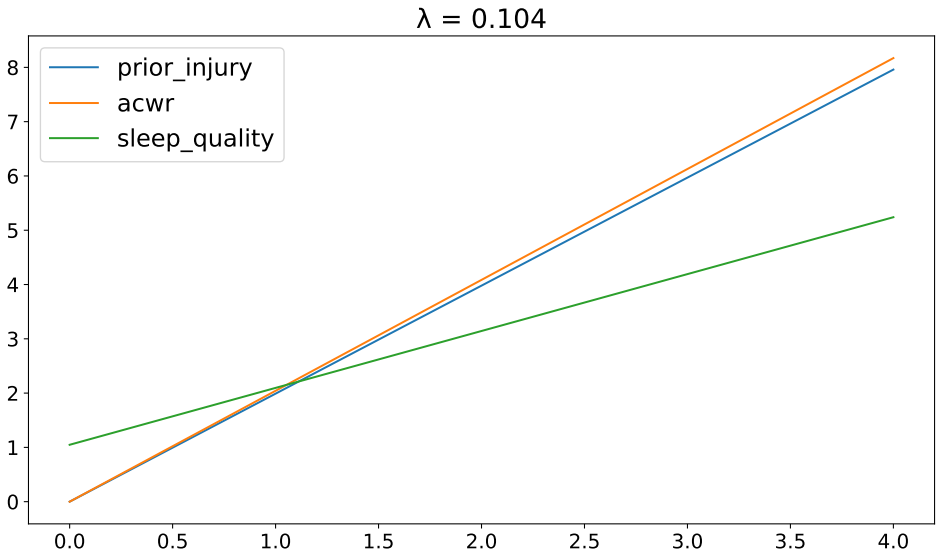
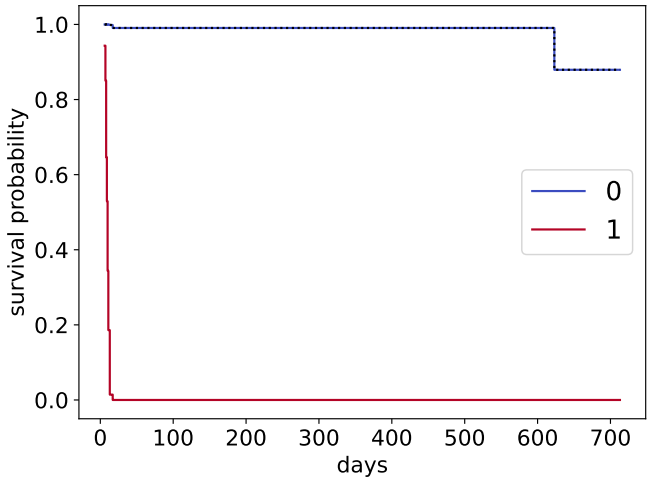
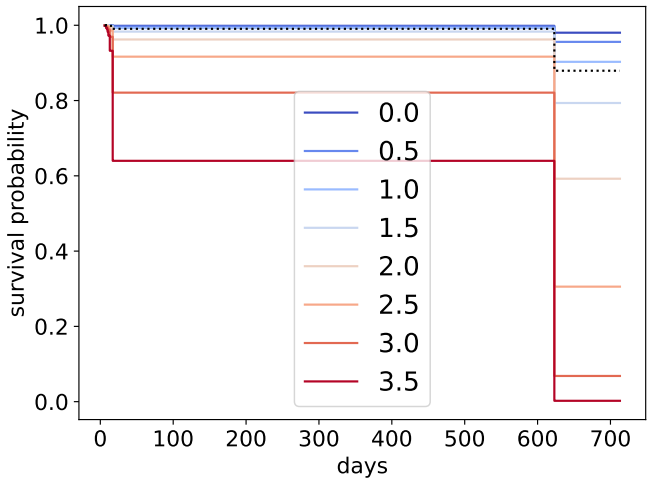


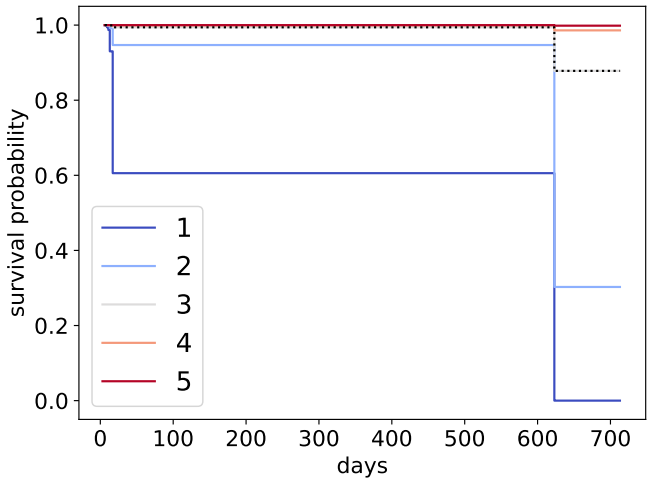

Figure A.4: Injury risk functions for optimal covariates using recurrent injuries from Team B




(a) Prior injury




(b) ACWR




(c) Sleep quality

Figure A.5: Partial effects on survival outcome of covariates using recurrent injuries from Team B