Analysis of National Hockey League Injuries

Sam Bolig, Jeffrey Ho, Antara Mondal, Aijin Wang

May 11, 2018

Abstract

While the National Hockey League (NHL) generates over 3 billion dollars in revenue annually, it is estimated that over 200 million dollars are wasted each year on paying injured players throughout the league. For this reason, players who are thought to be prone to injury are paid less and offered shorter contracts. However, no methodology for predicting a player's recovery time or injury duration exists. Furthermore, no published model can either predict injury recurrence or assess the impact of prior injuries on game performance. This study implements basic survival analysis, cox proportional hazard modeling and multi-state modeling in order to predict the probability that a player returns from an injury over time. Data was collected from The Sports Network's NHL website, which provides injury histories for all active NHL players. Using these observations and an adaptation of survival analysis, we model the likelihood that a player returns from a specific injury over time, given his characteristics. Using multi-state modeling, we also produce probability estimates for a player's likelihood of injury recurrence and recovery. These methods resulted in several models which all agreed that injury type is indeed a significant predictor of injury duration, as is a player's age and time spent in the NHL. However, a player's position, height and weight are not useful indicators of either his injury duration, or the likelihood of injury. This study aims to enhance NHL members' decision making processes based on players' injury histories.

1 Introduction

Each year, National Hockey League (NHL) owners pay out over 200 million dollars to injured (inactive) players [1]. This waste is not affordable if a team wants to remain competitive, as their annual spending is limited by a salary cap. For this reason, players "prone" to injury are held to a higher standard of play, and are offered shorter contracts for less money per year. Team doctors also must place limits on these players, such as minute restrictions, in order to extend their careers. These decisions ultimately affect the team's coaching and game plan, and can drastically affect their performance. In short, all parties in the NHL must consider what determines a player to be "prone to injury", as well as how to best predict their expected recovery time.

While there have been some epidemiological studies about sports injuries in general, such as *Issues in Estimating Risks and Rates in Sports Injury Research* [2], no prior work analyzes the NHL specifically. Furthermore, no studies yet attempt to predict injury recurrence, but instead recommend medical resources or equipment for the treatment of injuries. These papers do, however, utilize a few key terms that are quite relevant to the basis of our own study:

- Prevalence: The proportion of athletes who have an existing injury at any given point in time
- Incidence: The number of new injuries that occur over a specific period of time, such as from the start of the season.
- Incidence proportion: Average risk measure
- Clinical incidence: Used for resource utilization studies (e.g. clinicians want to know how many injuries they expect to treat in a season)

2 Objective

The purpose of this research has two aims: firstly, we attempt to model players' expected recovery times given a specific injury and other relevant predictors. Secondly, we propose a model for predicting injury recurrence in the NHL. This model should be able to produce probability estimates for the likelihood of injury at the player level. Both models would also report the significance of their predictors, which would provide insight into which factors affect players' health. These results have the potential to influence several parties' decision making in the NHL, such as: league owners' contracts, offers, and salaries, recruiters' talent scouting, coaches' game plan decisions, trainers' recovery programs, and each player's own understanding of their career. In conclusion, this research will introduce two formalized models for injury prediction in the NHL, which will provide information useful to all members of the league.

3 Data Description

Our data set contains the injury histories of 567 active players from all 31 teams in the NHL, each having 20-23 players on their roster. These histories date back as far as the year 2000, and are updated as recently as the current 2018 season. This results in a data table of 3950 injury observations, with each injury case as a single entry. For each injury, we have information on the injury type, the player injured, the time frame of the injury and other relevant descriptors.

The data for this project was collected from The Sports Network NHL website, specifically their "Player Bio" pages [3]. Within these web pages are tables listing all current players' injury histories, stored in JSON format. R, the statistical software utilized for this analysis, contains a package that allows JSON data to be read into an R table. The data can then be organized and cleaned in R itself for further analysis.

Each player's injury table contains information on their games missed and injury type. Because these descriptions are stored as character strings, parsing them for quantitative measures requires intensive manipulation. For example, injury dates were converted from qualitative strings to the number of days past since the start of the year. This quantity could then be used to measure the duration of players' injuries in a consistent manner. Furthermore, secondary measures of player characteristics were collected in the same manner from the from TSN website, which could then be used as covariate predictors.

3.1 Procedure

The data we collected has been scraped directly from the JSON tables of all current NHL players on the TSN's Player Bio web-page. We obtained this data by creating a Python Script that utilized Selenium, a web scraping tool. The player bio tables extracted contain information regarding transaction history, suspensions, fines by the NHL, and injury updates. After downloading and scraping each player's NHL transaction table, we could aggregate our data into a uniform table where each row corresponds to an injury event, and its columns describe that event.

The listings of injuries are semi-standardized. When an injury takes place, a player's bio page is updated with a row containing the injury type and date; the subsequent row then states the number of games the player missed once he returns to his team, with an updated classification of his injury type. However, there exists many cases in which the injury descriptions are updated after the initial event, and additional rows are included in the Bio table to reflect this. It is therefore crucial to not only record the initial date of injury and the date of return, but to also update the injury type with the most accurate description.

To process and clean players' injury data, we parse through the string output of each injury description. In this manner, our team has successfully collected each injury's updated type, the number of games missed due to each injury, and the days elapsed during the injury. Players' names, affiliated team information, and other characteristics were then attached to each injury observation.

3.2 Measures

The final dataset included 40 variables scraped from the TSN website. These 40 variables included basic demographic information of the player and their affiliated team, as well as information about their injuries, date of occurrence, how long they were healthy for before becoming injured, and whether or not they have previously been injured.

There were originally 273 distinct injury types. These injuries were then reduced 12 injury types: arm/hand, concussion, foot, head/neck/face, illness, leg, lower body, mid body, mouth, upper body, undisclosed, and other.

The models were created using the number of games injured, number of days missed, number of games missed, occurrence of previous injury, the start and end dates of the injury, and position covariates.

4 Exploratory Data Analysis

Figure 1 displays a conditional barplot of frequencies of injury locations by team. The proportion of injury types are fairly similar across teams. However, the count of certain teams, such as the Rangers and Blackhawks, have fewer injuries than the other teams.

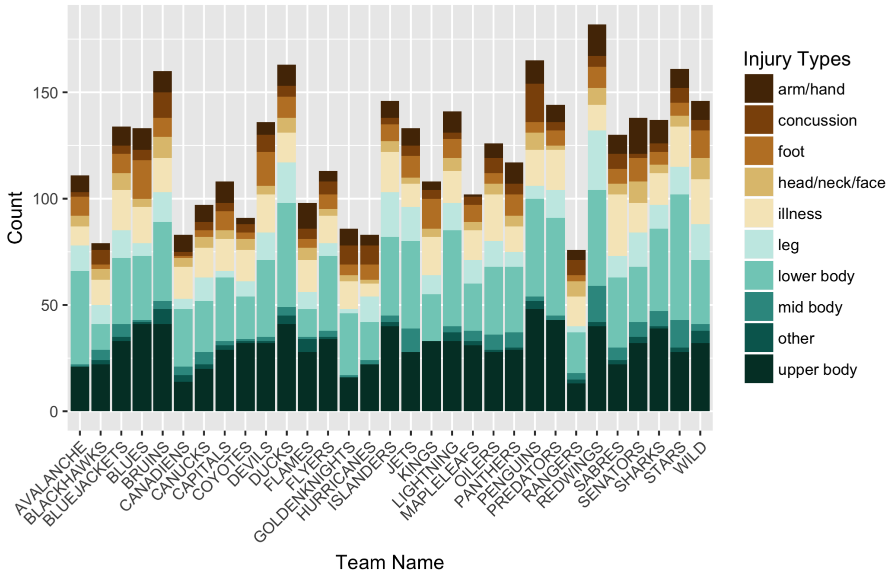


Figure 1: Frequency of Injury Types by Team

Figure 2 shows the distribution of the 'games missed' variable. It is severely right skewed, and the majority of the data is located between 0 and 14 missed games.

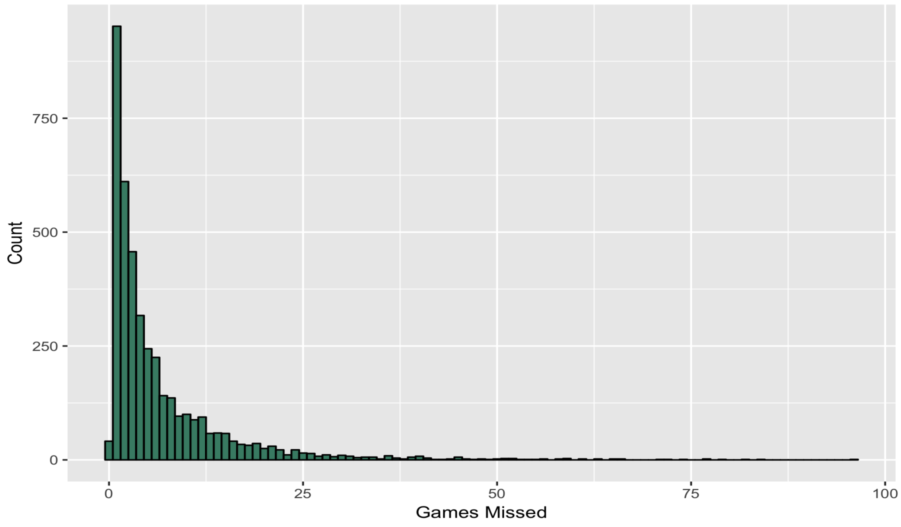


Figure 2: Distribution of Games Missed

Figure 3 analyzes the past two years of the Pittsburgh Penguins' injuries. This is to visualize what injuries look like for a typical team.

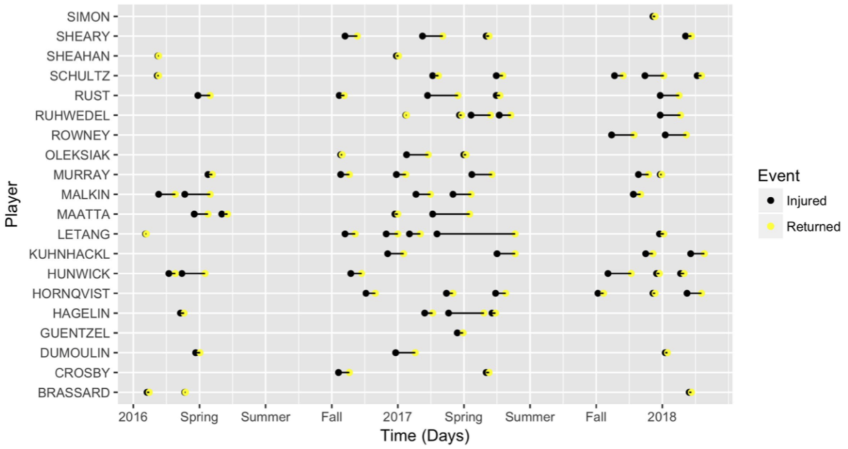


Figure 3: Injury Histories of the Pittsburgh Penguins (2016-2018)

Figure 4 analyzes any correlations between the continuous variables. We see that weight and height have high positive correlations, which may suggest use of interaction terms in modeling.

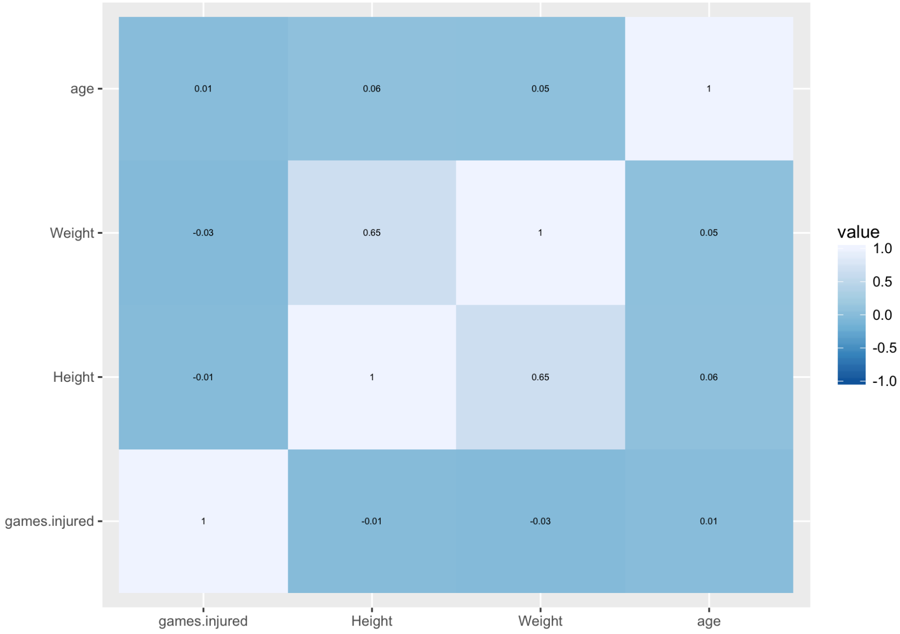


Figure 4: Correlation Heatmap of Numeric Covariates

5 Methods

5.1 Introduction

In this section, we will introduce three models that are related to the survival analysis: basic survival analysis, cox proportional hazard model, and multi-state model. In addition, we will compare and contrast the model fit among three models and look for potential patterns related to injury statistics.

5.2 Basic Survival Analysis

To estimate the probability that a player returns from his injury over time, we utilize Survival Analysis (SA) modeling. SA methods analyze the expected time until a given event takes place. SA derives its name from the medical field application, where SA models predict the probability that a patient will survive over time, given various treatments. Our study instead estimates the probability that a player will remain injured over time, given a set of variable inputs.

Survival Analysis relies on a specific data structure. The first component is a fixed starting point for each observation period. The probability of the event is then modeled as a function of time and treatment. In the context of our study, the starting point is the first day of a player's injury, the event is the player's return to the league, and his "treatment" is his injury type.

For cases where the event does not take place within the observation window, SA modeling uses right-censoring to account for limited information. Right-censoring observations reduces the influence of those points on the model, as they are less informative than complete observations. In the typical SA context, patients who do not follow up with the study or those who do not die within the observation window are right-censored. In our study's context, players who do not return to the league before the end of a given season are right-censored. For those players, a precise recovery time is unknown.

SA relies on the assumption of "non-informative censoring" which asserts that the censoring of data points is independent of the event taking place [4]. This implies that injuries with no return date are not systematically different than complete injury observations. This logically holds in our data, as the probability of a player's injury lasting into the summer does not relate to the probability that a player returns from his injury.

In the following sections, we will discuss the survival function for players with different type of injuries, as estimated by the Kaplan-Meier estimator. This is the most basic form of survival analysis we address. The Kaplan-Meier Curve formula is as follows:

$$\hat{S}(t) = \prod_{i:t_i \leq t} (1 - \tfrac{d_i}{n_i})$$

with t_i at time when at least one event happened, d_i the number of events that happened at time t_i(in this case, the event is the player returns from injury) and n_i the individuals known to survive at time t_i(here is when they are still injured).

A major limitation of SA modeling is the predictor variable itself. Because SA originated from medical studies with qualitative treatments, these models do not account for quantitative predictors measured on a continuous scale. This makes it difficult to account for factors such as a player's age or size. Secondly, SA modeling does not account for time dependent covariate effects. This is because SA modeling is usually applied to controlled studies with random assignment, so controlling for covariates is not a common concern. Furthermore, SA assumes independence of the events, but in our data, most of the players get injured again, whether or not due to previous injuries. We therefore employ a semi-parametric SA model, the Cox Proportional Hazards model, which has the ability to account for both quantitative inputs and time dependent covariates.

5.3 Cox Proportional Hazards Model

The Cox Proportional Hazards (Cox PH) model is a semi-parametric form of SA modeling, which takes in a greater variety of inputs (covariates) and measures their effect on a survival-object response variable in the form of hazard curves [5]. A survival-object response variable contains the time the event took place, or indicates the time elapsed until the end of the observation window (denoting these observations as right-censored). Inputs are entered into the Cox PH model linearly, and so their effects are coerced to be proportional over time. This allows for relatively simple interpretation of model coefficients; they report the expected change in log hazard ratio of an event taking place, given a unit change in the predictor and holding all else fixed. The log hazard ratio signifies the expected change in the likelihood of an event taking place. It is referred to as a hazard in the medical context, however in our model the event is a player's recovery date.

The Cox Proportional Hazards formula is as follows:

$$h(t) = h_0(t) * e^{b_1 x_1 + \ldots b_p x_p}$$

where t is again time, $h_0(t)$ is the baseline hazard where all covariates equal 0, and the $h(t)$ function then outputs a number which can be compared to the baseline. The b_i coefficients estimate predictors' effect on the hazard ratio in the following manner:

- $exp(b_i) = 1$: No difference in hazard from baseline
- $exp(b_i) > 1$: Increased risk of hazard
- $exp(b_i) < 1$: Decreased risk of hazard

The Cox PH model relies on the same non-informative censoring assumption as standard SA. It also assumes that covariate effects are proportional over time, hence its name. This assumption can be tested for post-modeling using adapted residual plots, which plot the predictor's estimated effect on the response variable over time. Finally, because SA modeling is usually applied to controlled experiments, individual observations are assumed to be independent and identically distributed. Because our study leverages injury observations from the same players over time, this independence assumption does not hold. Instead, to help account for these additional influences on the response variable, we include both time dependent (i.e. player's age) and time invariant predictors (i.e. player's height) in our Cox PH model. Including these inputs helps isolate the individual effects of each predictor. Our final model will then predict players' recovery times, and report the effect and significance of its predictors. We can then resample and refit these models using a boot-strap process to assess the variability of these estimates given new data.

5.4 Multi-State Modeling

The most common approach to analyzing survival data is the Cox proportional hazard model, as explained above. However, this model assumes independent observations, and it disregards information outside of a set observation window. Therefore, we also analyze our data using the multi-state model. This model is particularly useful in our study because we are able to determine how long a player takes to go from healthy to an injured state based on a specific injury and on previous information specific to each player.

Multi-state models can model the transition between several possible states that an individual can be in. This is beneficial for modeling various events that may be dependent on one another. Multi-state modeling is also useful for modeling recurrent events. The traditional survival analysis approach automatically creates two states: alive and death. In our study, we focus on the transition between a player's healthy and injured state. With multi-state modeling, we can model the transition between states, or an event.

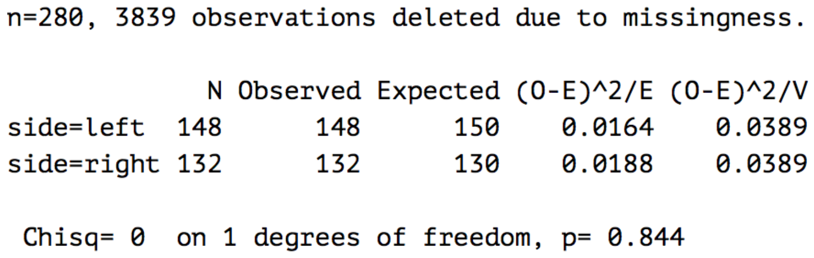
To estimate a multi-state model, we require a state variable and time variable at the least. The state variable, in our case, is a specific injury or the 'injured' state, and the time variable is the number of years since 2000. Each subject has a unique ID identifier. A more advanced model consists of multiple covariates that the function will take into account. The model returns a transition matrix, indicating the rate at which players switch between states or remain in the same state.

6 Results

6.1 Basic Survival Analysis

In the earlier discussion, we stated that over the years, players tend to get injured again. From the data set and exploratory data analysis, we also observe that the recovery time differs depending on if the players got injured before. Therefore, we make the speculate that whether players were previously injured may affect their recovery time at present. For the survival analysis, two variables are specifically explored: the injury location and whether the player is previously injured. We then try various criteria to categorize injury locations to identify possible patterns. The function $survdiff$ from package $Survival$ is used to test whether the two survival curves are identical.

First, when the injuries are categorized by left and right part of the body, the output of the function is as follows:



```
n=280, 3839 observations deleted due to missingness.

             N Observed Expected (O-E)^2/E (O-E)^2/V
side=left  148      148      150    0.0164    0.0389
side=right 132      132      130    0.0188    0.0389

 Chisq= 0  on 1 degrees of freedom, p= 0.844
```

The $p - value$ of the log-rank test is 0.844, suggesting that there is not enough evidence to prove that the time it takes to recover for left vs. right side injuries is statistically different. See below for the survival plot. 50 percent of the players were recovered in about 10 games for both left and right sided injuries.

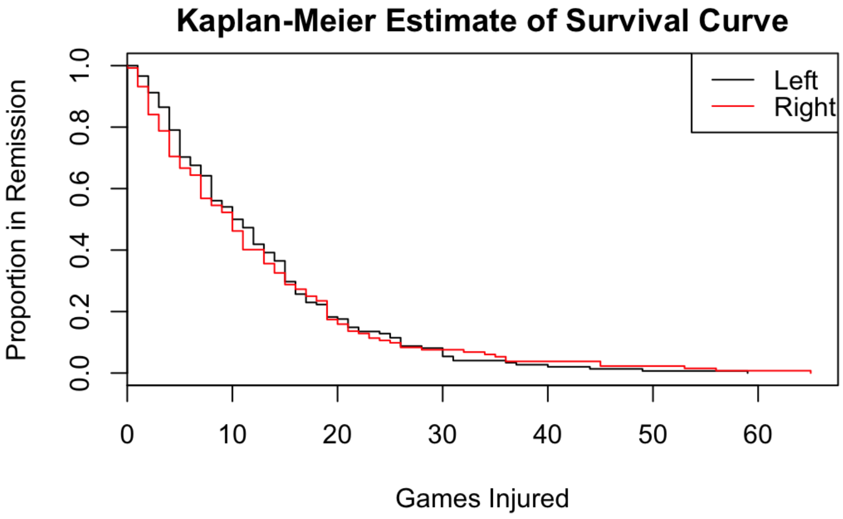


Figure 5: Survival Curves of the Recovery Time given Side of Injury

Let's take a deeper look into the possible association between $previouslyinjured$ with side of the injury. Figure 6 suggests that the two variables are not correlated.

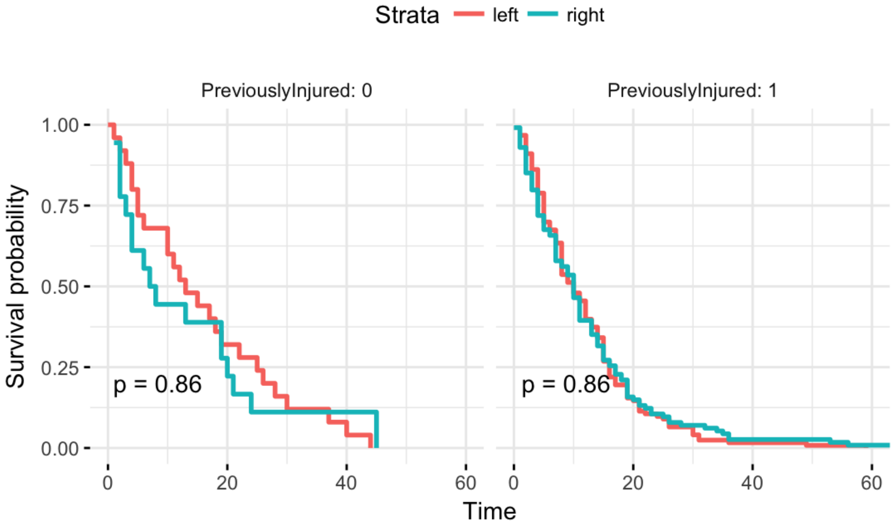


Figure 6: Survival Analysis of the Recovery Time given Side of Injury and Previously Injured

The reason that the two survival curves look so different for $PreviouslyInjured = 0$ but the $p - value$ is so big might be due to the small amount of data points. Since not all of the injuries have left or right indicator, we are only able to categorize 400 data points using this criterion.

Similarly, the analysis is performed on the body location of the injuries (upper vs. lower). Different from the previous criterion, the $p - value$ of upper vs. lower is 0.0274, suggesting a significant difference between the two injury locations. Figure 7 shows that upper body injury has a faster recovery time than lower body in average, and the model results show that 50 percent of the players who had upper body injuries came back in 4 games, while for lower body injuries it is 5 games.

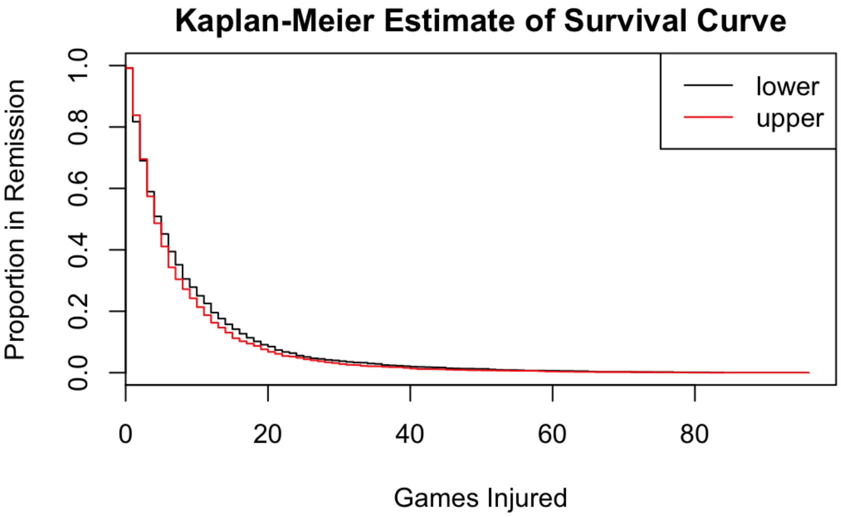


Figure 7: Survival Analysis of the Recovery Time given Body Location of Injury

What's more, The left graph in Figure 8 shows that for players who were previously injured, an upper body injury has a significant faster recovery time than lower body injury, while the right graph shows that if the players don't have previous injuries, the difference is insignificant.

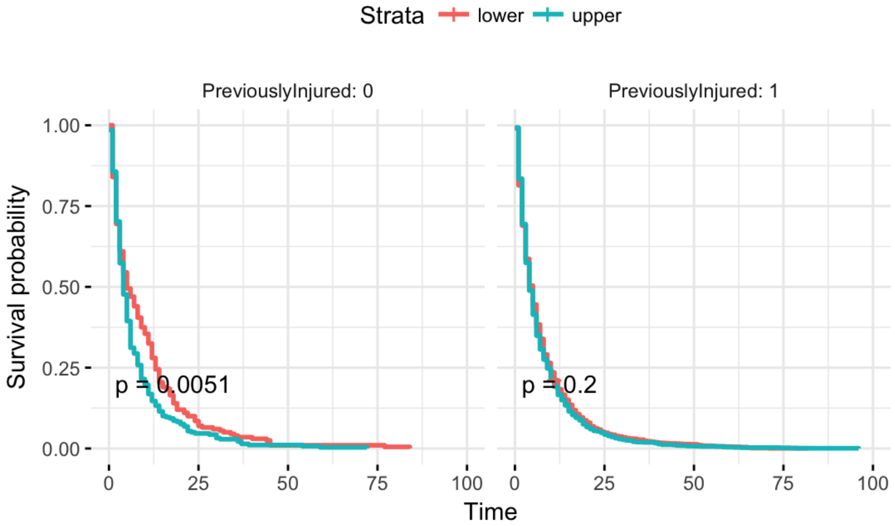


Figure 8: Survival Curves of the Recovery Time given Body Location of Injury and Previously Injured

6.2 Cox Proportional Hazards model results

Firstly, it is important to note that a portion of players in our data table did not have draft years recorded, so our data cleaning process did not record a "years played in the league" variable for these cases. The Cox PH model does not deal with incomplete observations, and so these players were excluded from the model. This leaves 3613 observations in our training data. The deletion of incomplete cases should not bias our model estimates, assuming that undrafted players' recovery times do not differ from others systematically. This assumption can be tested for in basic EDA plotting. Some examples of this testing are demonstrated in Figures 16-18 in the Appendix. These plots show that excluded cases do not differ significantly in injury type. However, the group of excluded observations tended to be slightly older on average, and their average injury duration was slightly less than for complete observations.

The Cox PH model's results are reported in Table 1. This model was created using R's 'Survival' package. The coefficient estimates listed report the change in hazard ratio given a unit change in the predictor after converting from the log scale. Their standard errors are listed on the log scale in the parenthesis, and their significance is denoted by the asterisks. If a predictor's coefficient is greater than one, it is associated with a greater probability of the event taking place at a given time. Because our event is a player's return to the league, it follows that larger coefficients correlate with shorter expected recovery times.

Several injury types are reported to be statistically significant, specifically: concussion, head/neck/face, illness, upper body and lower body. When an injury type is undisclosed it also is reported to be statistically significant. Each of these injury types have coefficients greater than one in Table 1, which means they all correspond with shorter expected recovery times on average, to varying degree. Figure 8 plots the expected recovery times for these selected injury types against all others, given a typical (having the mean value of each numeric predictor and the mode for each qualitative predictor) NHL player. It is consistent with the results in Table 1, as the other category has the longest expected recovery time and each predictor group lies below that curve.

A player's age and years played in the league are also both reported to be statistically significant predictors in Table 1, and age has a coefficient slightly greater than one while years played is slightly below one. We therefore found that older players are actually predicted to have shorter recovery times, but those who spend more time in the NHL are estimated to have longer recovery periods. The presence of a previous injury, too, is statistically significant and has a coefficient greater than one. Therefore players with at least one prior injury are expected to have shorter recovery periods. Finally, only the "Defense" position is reported to be statistically significant at any level among the position inputs, and this predictor actually relates to extended recovery times on average.

6.3 Cox Proportional Hazards Model Residuals

The Cox PH model assumes that the relationship between its predictors and the response variable is a constant proportion over time. To test this assumption, we utilize a built in function in R's Survival package. This function assesses the proportional relationship for each predictor individually using a form of residual plotting; if a predictor's relationship with the response variable is proportional over time, its residual plot will have a slope of 0. An example of the residual plots output is produced in Figure 10. These plots show a predictor that appears to

Table 1: Cox PH Model Results

Dependent variable	
concussion	1.2330* (0.114)
foot	1.0835 (0.105)
head/neck/face	2.0390*** (0.122)
illness	7.6759*** (0.094)
leg	0.9461 (0.097)
lower body	1.8589*** (0.084)
mid body	1.0878 (0.124)
mouth	0.9342 (0.299)
other	1.0332 (0.179)
undisclosed	3.4666*** (0.104)
upper body	1.8858*** (0.085)
previously injured	1.0963* (0.055)
height	1.0076 (0.013)
weight	1.0015 (0.002)
Defense	0.9168* (0.049)
Goaltender	0.9253 (0.077)
Left Wing	1.0173 (0.056)
Right Wing	1.0193 (0.057)
age	1.0832*** (0.025)
years.played	0.9100*** (0.026)
Observations	3,613
R^2	0.215
Max. Possible R^2	1.000
Log Likelihood	−22,465.300
Wald Test	1,020.580*** (df = 20)
LR Test	873.062*** (df = 20)
Score (Logrank) Test	1,196.921*** (df = 20)
Note:	*p<0.1; **p<0.05; ***p<0.01

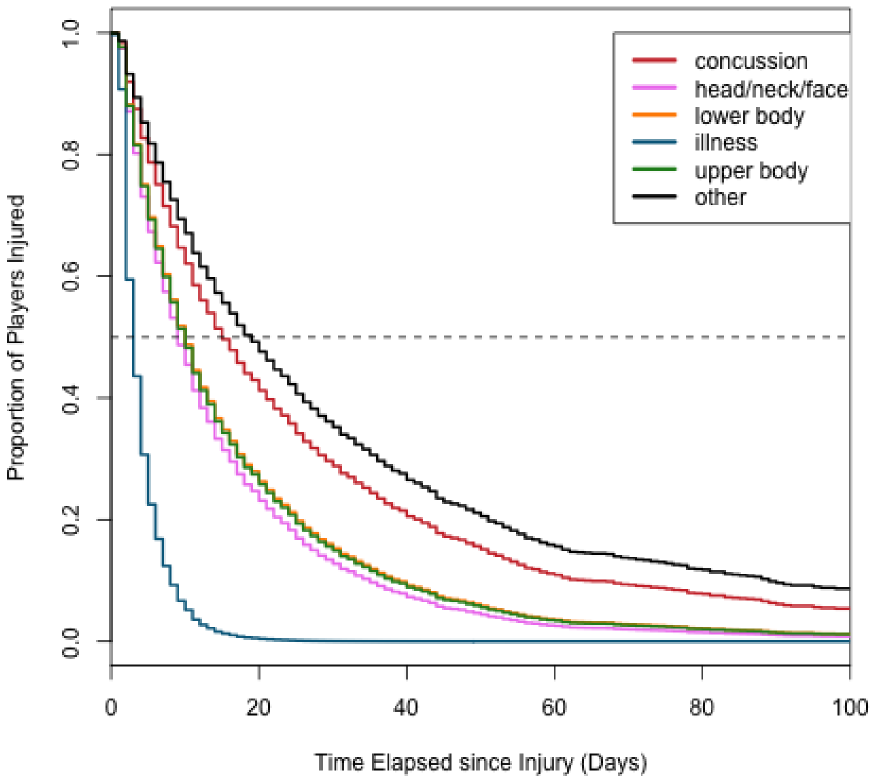


Figure 9: Cox PH Model Curves given Injury Type

satisfy this assumption alongside three predictors with features that break this assumption. The exceptions plotted in Figure 10 are for the illness, undisclosed and upper body predictors which all show slight deviation from the slope 0 line. Figures 19-22 in the appendix demonstrate that the majority of our predictors have roughly constant relationships with players' hazard ratios over time, however deviation is often present in injuries with longer duration. To account for this deviation, the Cox PH model can include an interaction term between a predictor and time, which helps to model the dynamic relationship between these variables. Another interesting feature of these residual plots is the distinct groupings along the y-axis, which again measure the predictor's effect on the hazard ratio. These groupings indicate that our model may not be accounting for all relevant predictors, as some latent variable may be influencing the model to over/under estimate particular observations in this consistent pattern.

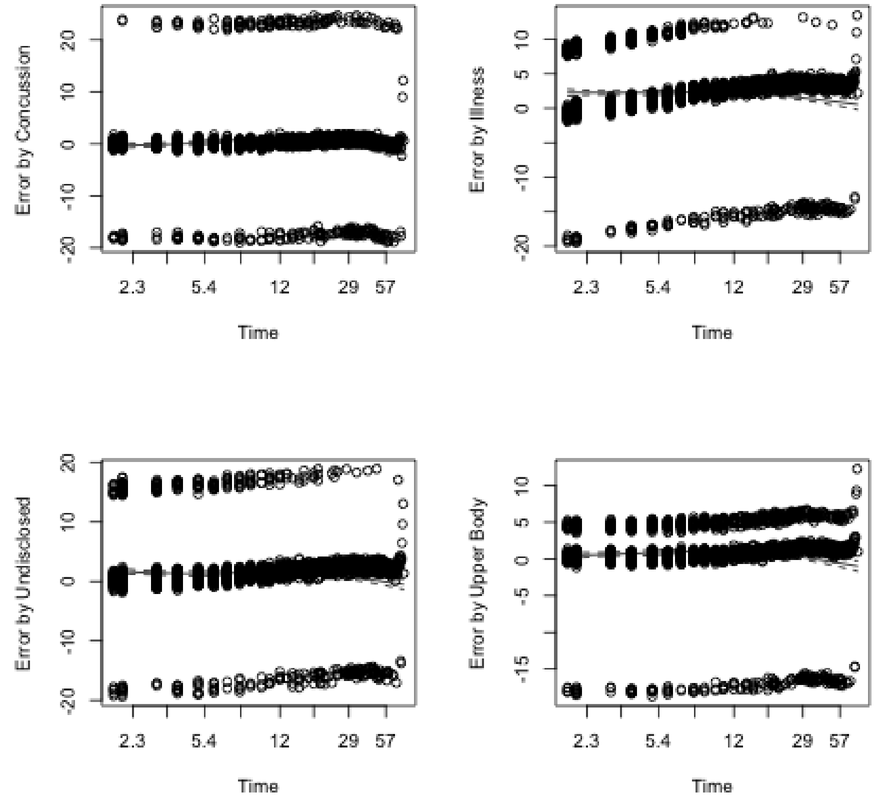


Figure 10: Cox PH Model Residuals assessing Proportionality

6.4 Cox Proportional Hazards Model Variability

To assess the consistency of our model estimates, we should not rely on the standard errors produced in the Cox PH model output alone. These standard errors are skewed because they rely on the same training data that was used to create the original estimates. To better understand the variability of these coefficient estimates, we will employ non-parametric boot-strapping. By resampling the original training data and producing new model estimates, we can estimate the consistency of our model's coefficient estimates. We performed this boot-strapping 1000 times, and then obtained 1000 p-values for the original Cox PH model's significant predictors. These p-values are plotted in Figures 10-12 below. Vertical dashed lines indicate the typical significance levels (0.01, 0.05, and 0.1). We can see that almost all of the p-values generated using this boot-strapping technique are statistically significant. Three of the more variable coefficient estimates are the concussion injury type, the previous injury dummy variable, and the Defense dummy variable. This is logical, as all of these estimates had larger standard error estimates in the original Cox PH model output. Overall, the consistency of these coefficient estimates and their statistical significance is extremely promising.

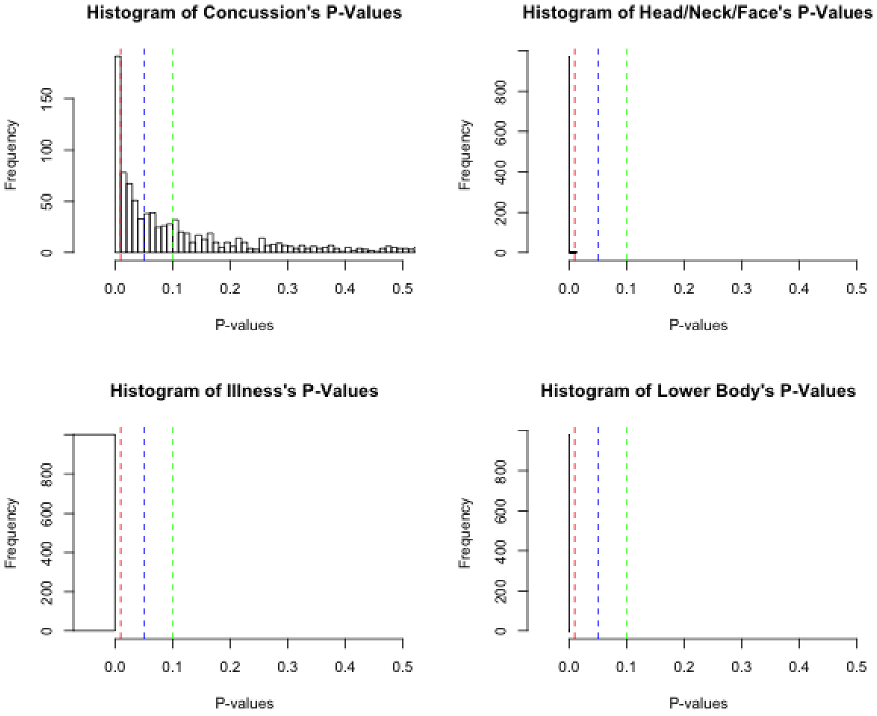


Figure 11: Cox PH Model Boot-strapped P-values

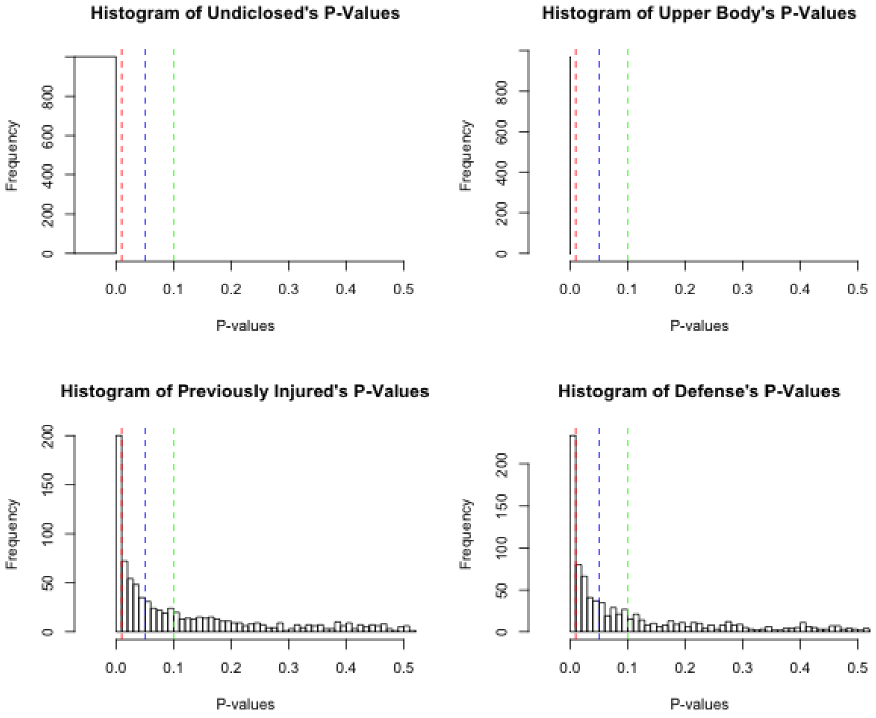


Figure 12: Cox PH Model Boot-strapped P-values

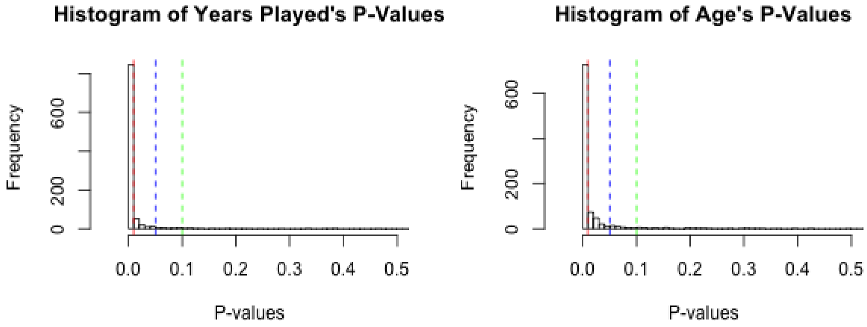


Figure 13: Cox PH Model Boot-strapped P-values

6.5 Multi-State Model

Various multi-state models were created with different combinations of our covariates. The null model modeled solely injury transition vs time without any covariates. This model outputs the following transition matrix.

$$Q = \begin{bmatrix} -q_{ih} & q_{ih} \\ q_{hi} & -q_{hi} \end{bmatrix} = \begin{bmatrix} -22.92 & 22.92 \\ 1.45 & -1.45 \end{bmatrix}$$

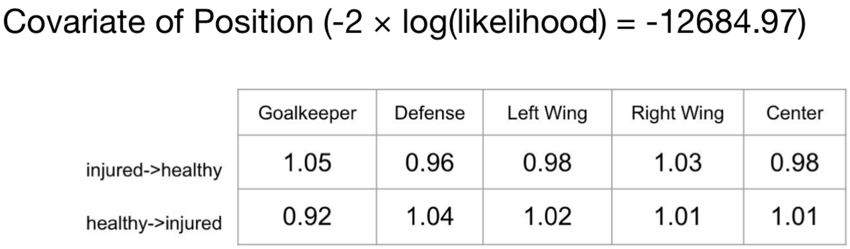
In this matrix, q_{ih} denotes the rate of transition between the injured and the healthy state. The time until a player recovers is exponentially distributed with rate q_{ih}.

Given that for X$\sim$Exp(λ), $\exp X = 1/(\lambda)$

- Injured – Healthy: (1/22.92) x 365 $\approx$ 16 days
- Healthy – Injured: (1/1.45) x 365 $\approx$ 252 days

The null model estimates using the exponential distribution that the expected number of days it takes for a player to go from an injured to a healthy state is about 16 days, and to go from a healthy to an injured state is about 252 days. The following two models add covariates to further create more accurate estimates, and find which covariates are significant in estimation.

The next model (Figure 18) included the ”position” variable. This variable had five levels: goalkeeper, defense, left wing, right wing, and center.

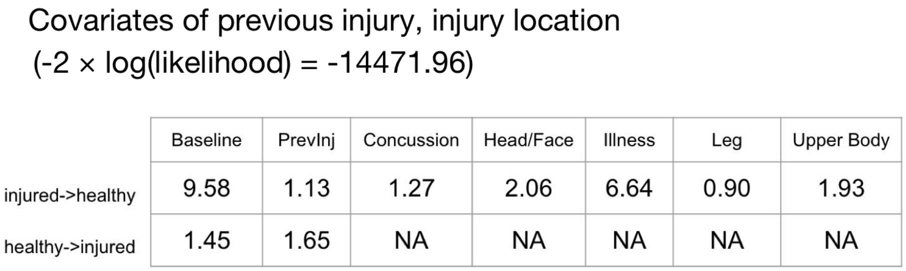


Covariate of Position (-2 × log(likelihood) = -12684.97)

	Goalkeeper	Defense	Left Wing	Right Wing	Center
injured->healthy	1.05	0.96	0.98	1.03	0.98
healthy->injured	0.92	1.04	1.02	1.01	1.01

Figure 14: Position Transition Matrix

The next model included the binary variable previous injury and the injury location (Figure 19).



Covariates of previous injury, injury location
(-2 × log(likelihood) = -14471.96)

	Baseline	PrevInj	Concussion	Head/Face	Illness	Leg	Upper Body
injured->healthy	9.58	1.13	1.27	2.06	6.64	0.90	1.93
healthy->injured	1.45	1.65	NA	NA	NA	NA	NA

Figure 15: Position Transition Matrix

To compare the various multi-state models, likelihood ratio tests were conudcted to determine the significance of covariates that were tested. The first likelihood test that was run tests the significance of a player's position on injury recovery times and the second likelihood test that was run tests the significance of any previous injury on recovery time, as well as the location of a given injury.

```
# Likelihood Ratio Testing for Covariate of Player Position
> lrtest.msm(msm.null, msm.pos)
        -2 log LR df         p
msm.inj  6.426282 10 0.7782684
> lrtest.msm(msm.null, msm.previnjured.injurytype)
                       -2 log LR df p
msm.previnjured.injury  1202.064 14 0
```

For reference, the position variable contains five categorical values: Goalkeeper, Defense, Left Wing, Right Wing, and Center. For the purposes of keeping consistent categorization, the "Left Wing" and "Right Wing" categories were not combined in significance tests with position as a covariate, despite relatively similar playing styles between the two positions. The injury location variable tells us which types of injuries keep players out for longer periods of time.

As seen above, at a p-value of 0.778, the model conditioned on a player's position was found to not be statistically significant in terms of its contribution to transitioning from an injured state to a healthy state and vice-versa. However, at a significance level of $< 2 \times 10^{-16}$, the model that conditions a player's current state on previous injury and location of a player's current injury is statistically significant. Throughout multiple tests of significance in comparing log-likelihood values of different models, we ultimately found that injury location and any previous injury history to be the most statistically significant covariates for multi-state modeling.

7 Conclusion & Discussion

All three of our modeling techniques produced consistent trends in estimating predictors' relationships with recovery time. The basic survival analysis shows that survival curves appear to depend on injury types, and specifically, differences are found between upper and lower body injuries, especially if the players were previously injured, but none for left and right body injuries under any circumstance.

In the adapted Cox PH Hazard model, most injury types again report statistically significant coefficients. Furthermore, players' age and longevity appear relevant to estimating recovery time, as well as the presence of a previous injury. Surprisingly, a player's position is apparently not too relevant a factor in modeling his recovery timetable. These results held in repeated resampling of the training data, signifying the pattern is consistent given new data. The Cox PH model results have slight bias due to the deletion of incomplete data, however the direction of this bias is predictable; because the omitted data tended to come from older players with shorter injury duration, the training data would associate older players with longer injury duration. This implies the Cox PH model results may report age to be more relevant a factor than in reality. Overall, the Cox PH model estimates players' expected recovery times.

The multi-state model's results, too, are useful for teams to gauge how long it will take a player to recover based on a specific injury. Additionally, coaches and sports physicians may be able to take further steps in preventing/avoiding injuries that have longer recovery periods.

Individual players can also attempt to prevent further injuries from occurring by understanding how their injury histories relate to their recovery periods.

7.1 Limitations

One limitation we faced was that after 2007, injury reporting became highly generalized, such as "upper body injury" or "lower body injury". Thus, our injury types have become more generalized in recent years. This was out of our control, however, this is understandable because teams may not want their competition to know which types of injuries their players are prone to.

Furthermore, we only fit models using the existing injury data from TSN, but don't take human factors into account. For instance, in reality, players may lie about their injury severity in order to participate in more games. In this case, the recovery time for some injuries may not be accurate. This trend seems fairly consistent with the lack of specificity in more recent injury reportings.

7.2 Future Work

One possible action we can take is to work closely with teams to acquire more accurate injury data. In addition, we can analyze recovery time and recurrence patterns for each type of injury.

8 Appendix

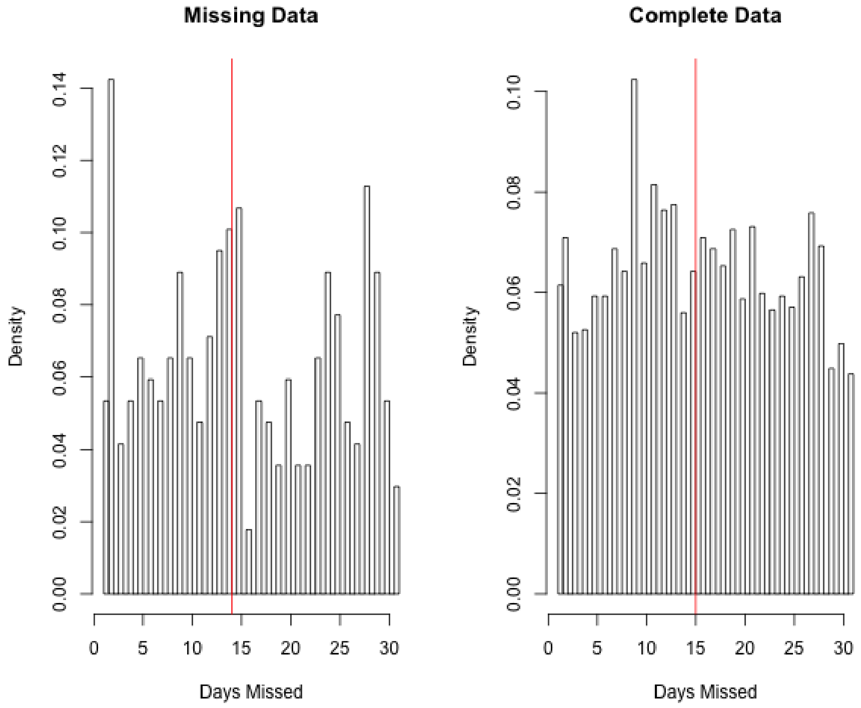


Figure 16: Comparing Complete and Incomplete Data Sets: Days Missed Dist.

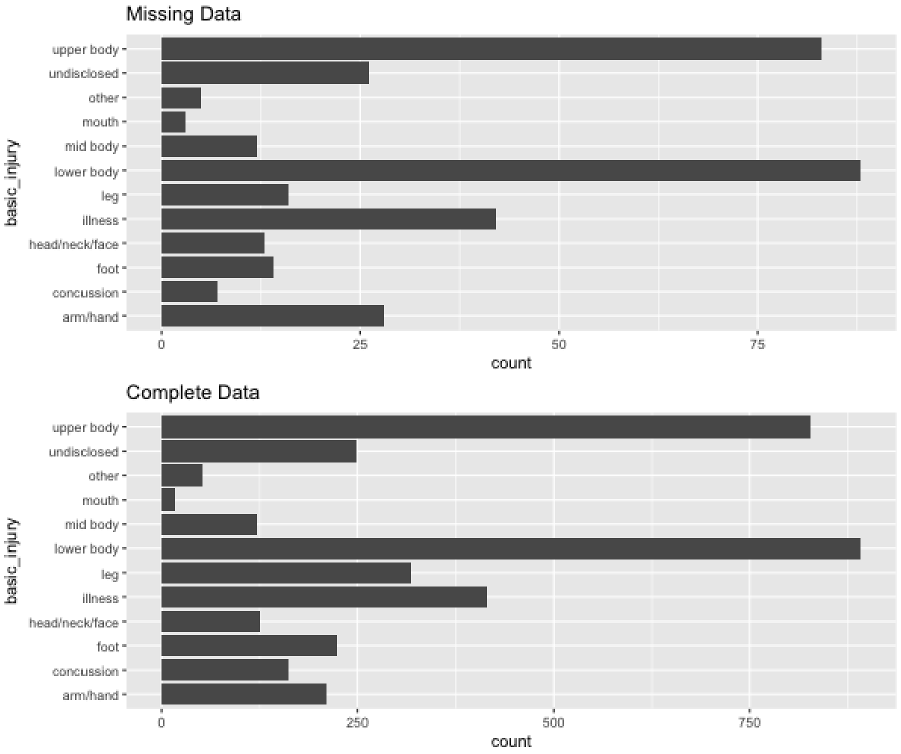


Figure 17: Comparing Complete and Incomplete Data Sets: Injury Dist.

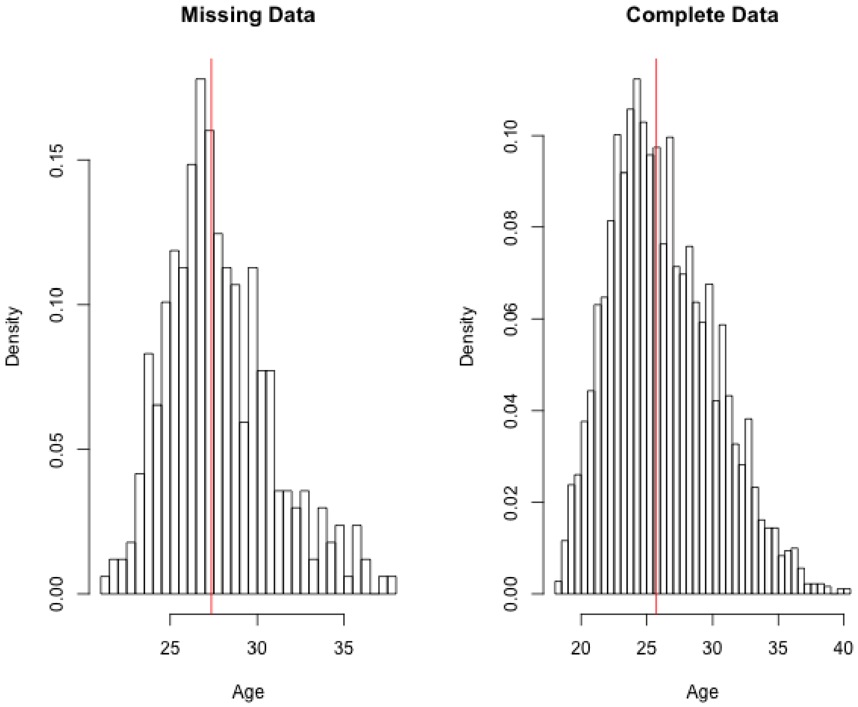


Figure 18: Comparing Complete and Incomplete Data Sets: Age Dist.

9 References

1. Caba, Justin, "NHL Owners Have Paid Out Over 650 Million In The Last 3 Seasons To Injured Players Who Were Not Competing" (Medical Daily).

2. Knowles SB, Marshall SW, Guskiewicz KM. Issues in Estimating Risks and Rates in Sports Injury Research. Journal of Athletic Training. 2006;41(2):207-215.

3. https://www.tsn.ca/nhl/players

4. Ranganathan, P., & Pramesh, C. S. (2012). Censoring in survival analysis: Potential for bias. Perspectives in Clinical Research, 3(1), 40. http://doi.org/10.4103/2229-3485.92307

5. Fox & Weisberg, (2011). Cox Proportional-Hazards Regression for Survival Data in R. https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Cox-Regression.pdf

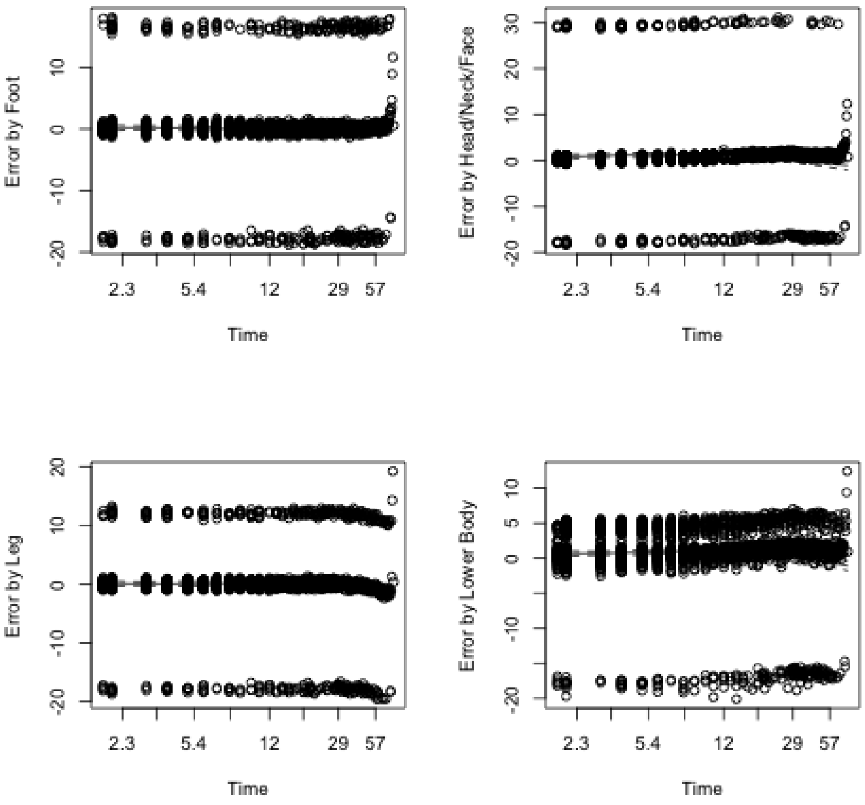


Figure 19: (More) Cox PH Model Residuals assessing Proportionality

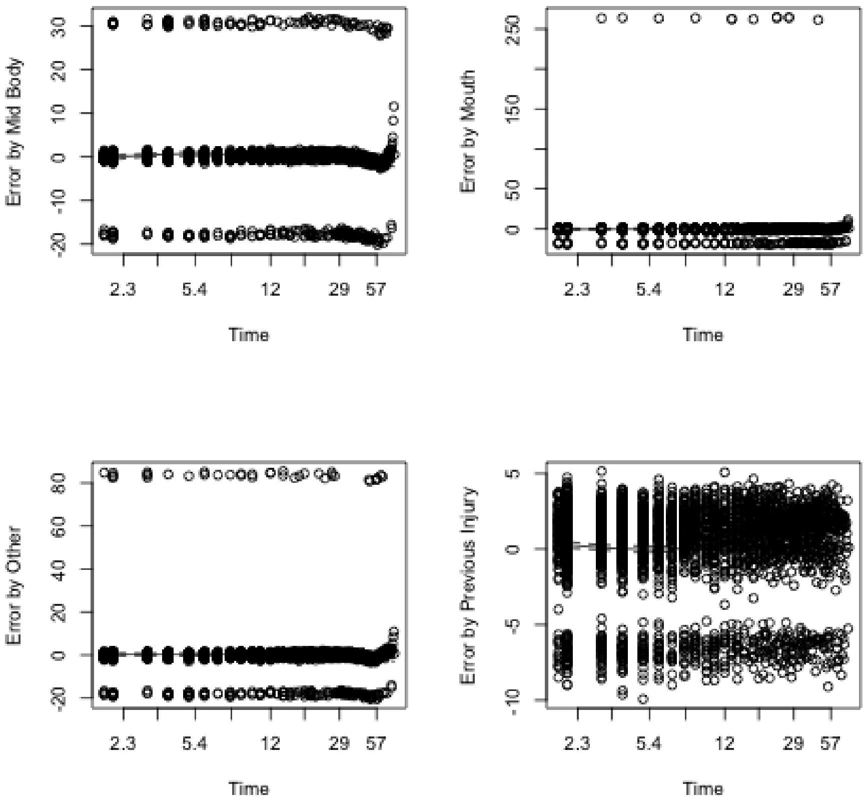


Figure 20: (More) Cox PH Model Residuals assessing Proportionality

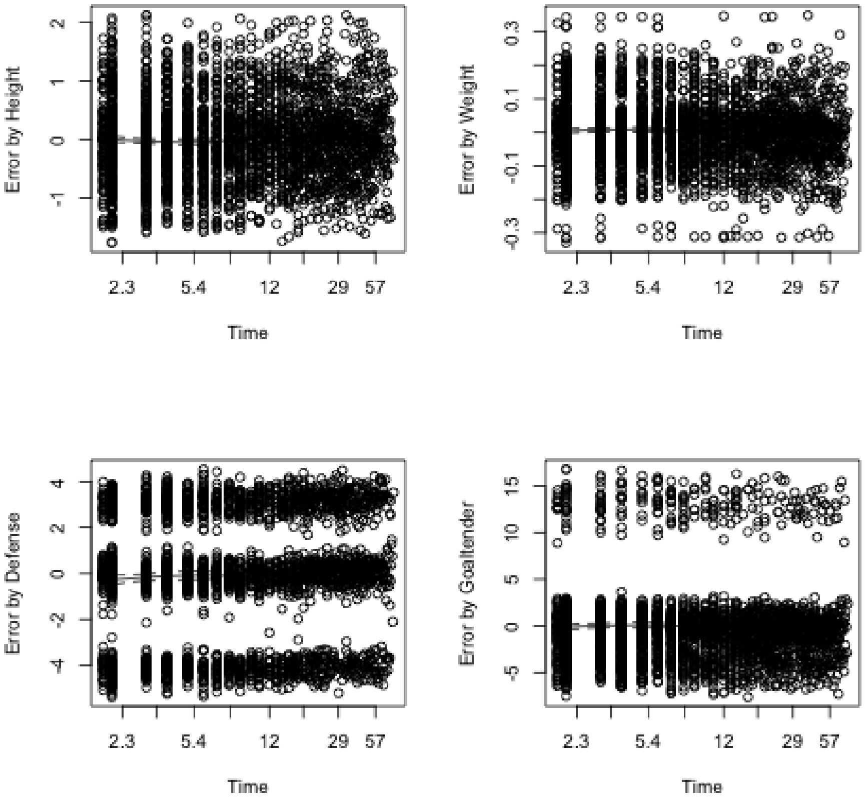


Figure 21: (More) Cox PH Model Residuals assessing Proportionality

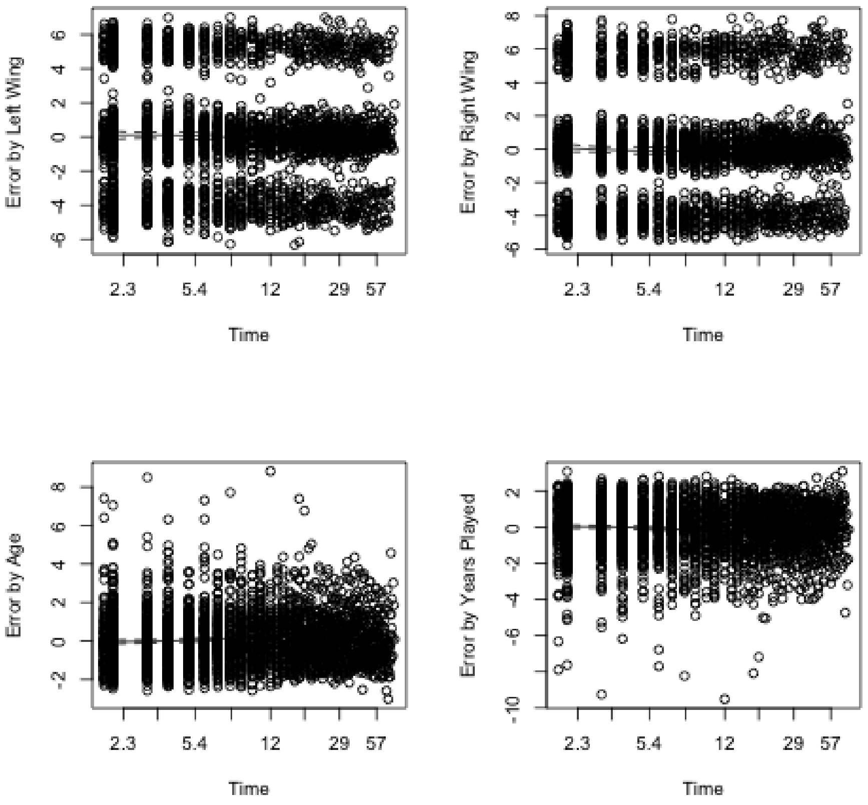


Figure 22: (More) Cox PH Model Residuals assessing Proportionality

10 Contributions

- Sam: My major contribution to this research project involved the Cox PH modeling. To accomplish this, I first re-formatted our data to fit the model criteria. I then ran multiple iterations of this modeling, using covariates in the data as well as manually created covariates (such as the years played variable). I assessed the model using residual diagnostics and boot-strapping. Finally, I analyzed possible issues with the model creation, and attempted to explain its results in the context of our larger objective. Before we even began modeling, I helped to explore the data we collected, and proposed possible explanatory variables of interest. More generally, I drafted and revised the opening sections of this research paper, and I helped delegate roles within our team. I was consistent in my communication throughout the semester both within our team and with our advisor, so that we understood our progress and goals at all stages.
- Jeffrey: I worked on two main components of this research project: data processing and multi-state modeling. To scrape data, I learned the Selenium Web Scraping framework in Python to navigate to the correct pages and extract JSON components. Throughout the data cleaning process, I spent much of my time at the start by ensuring duplicate entries and inconsistencies were addressed. On multiple occasions, I went back to scrape additional datasets in order to analyze more covariates. In addition, I did much of the MSM analysis by building our final models and running significance tests on them. This constituted into one of the three main models of our final research.
- Antara: During the beginning stages of the project, I did some of the EDA to gain a sense of understanding of the data. I also looked up any previous research that was completed to gain an idea of what is typically studied in relation to sports injuries. After we began the modeling stage, I worked mainly on learning the theory behind and implementation of MSM modeling, and figuring out how it applied to our data. I fitted multiple MSM models to understand if any of our covariates were significant by testing multiple injury types. In addition, I drafted our initial abstract and final MSM modeling section of the paper. More recently, I went through many of the EDA and MSM edits suggested, such as editing our graphs and adding additional information.
- Aijin: I did two parts in this research. After we got the data and did the initial cleaning, the entries weren't consistent, so I spent sometime determining edge cases using different criteria. Afterwards, the data format didn't match with the models that we wanted to implement. For the basic survival analysis, we have to have season-ending variables, which were relatively hard to calculate directly from the existing data set(this part actually took a very long time). So I went through the data and explored different ways to mutate and clean the data further more to get the right variables set up. After spring break, I was mainly working on fitting basic survival analysis to multiple variables and identifying potential patterns. I did some EDA plots and drafted the poster as well.