NHL Injury Analysis

Sam Bolig, Jeffrey Ho, Antara Mondal, Aijin Wang

March 2018

1 Abstract

1.1

WRITE AT END

2 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 then offered shorter contracts and less money per year. Team doctors also must place limits on these "injury prone" 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 chemistry and performance. In short, all parties in the NHL must consider what determines a player to be "prone to injury".

There have not been any studies completed about hockey, specifically. There have been some epidemiological studies about sports injuries in general, such as *Issues in Estimating Risks and Rates in Sports Injury Research* [2]. These studies do not predict injury recurrence, but rather recommend medical resources or equipment. From these papers, there were a few key terms that were consistently used, which fit the overall basis of our study.

Some notable key terms are:

- 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)

To date, no study has proposed a formal method for determining 'injury proneness'. The purpose of this research is to propose a model for predicting injury recurrence in the NHL. This model should be able to produce probability estimates for the likelihood of injury given specific players' inputs. This model would also report the relative significance of certain measures' relationship with injury recurrence. Finally, as a secondary objective, further modeling will assess the impact of "injury proneness" on a player's performance in the league. These results will influence: 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 a formalized model for injury prediction in the NHL, which will provide information useful to all members of the league.

3 Data Description

3.1 Summary

The data for this project will mainly be pulled from the Toronto Sports Network (TSN) website, specifically their "Player Bio" pages [3]. Within these web pages are tables listing each player's injury histories, stored in JSON format. R 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.

Secondary measures on player performance have been pulled from the internet as well, either from TSN's website again or from NHL.com's "Player

Statistics" pages.

There are 31 teams in the NHL, each with 20-23 active players. Our sample therefore is comprised of data of 579 individuals.

Each player's injury table contains information on their games missed and injury type. It is stored as a character string, so parsing it for quantitative measures will require some manipulation; for example, we currently plan to create dummy variables for injury type.

3.2 Procedure

The data we collected has been scraped directly from bio tables of all current NHL players on Canadian-based sports network (TSN - The Sports Network). We obtained individual training data by creating a Python Script that utilized Selenium, a web scraping tool. The player bio tables we extracted contain information regarding transaction history, suspensions, fines by the NHL, and injury updates. After downloading each player's NHL transaction tables, we could aggregate our data to include the team a player was on during a given year, the exact time of injury, and other quantitative descriptors regarding their injuries.

The listings of injuries are semi standardized - for those with accurate listings, an injury type is listed in one row when the player's injury is first reported; on the subsequent row, it states the number of games the player missed. However, due to misclassification or specification of injuries later in time, there exists many cases in which the injury data contains an extra entry for an injury that specifies the injury location in more detail.

The process of cleaning our data involves being able to parse through the string output of a given injury description. From the data, we have been able to parse the type of injury, the number of games missed, and players' affiliated team information. In addition, using R's *dplyr* package, we were able to remove duplicate rows by checking for specific string patterns in subsequent lines of data.

3.3 Measures (UPDATED)

Our finalized data set contains 3,762 observations of injuries from 30 teams and 579 players in the NHL. Each observation then contains 7 descriptive variables. These include: Return Date, Injury (Type Duration), First Last Name, City Team, and Season Ending (T/F). **IN THE WORKS** We are also working on scraping the original Injury Date, in order to better model the player's time away from his team (opposed to modeling games missed only).

These variables were then expanded to create an additional 6 categories which are more applicable to statistical analysis; 'Date' was separated into 3 columns by YYYY / Mon / DD, 'Injury' was split into 2 columns for Type and Length separately, and in another column players were given unique IDs by combining 'First-Name.Last-Name.Team-Name'. **IN THE WORKS** Our additional Injury Date variable will also need to be split into a similar format, which will create an additional 3 columns, bringing the total to 15 at this stage.

Using these existing measures, another variable was artificially created indicating time elapsed between injuries for each player. To accomplish this, the date of each injury was first transformed into a column for 'Days elapsed by month', which was then added to the day of the month to create a variable named 'Day in the Year'. This variable was then subtracted from the previous injury (excluding each players first injury) to create a variable for days elapsed between injuries. **IN THE WORKS** A similar process will be used to identify the duration of each player's injury in days, which is the main variable of interest in our Survival Analysis model.

*** Injury Variable - Status Update***

Our original dataset came with 255 distinct injury types, but many were either duplicates or extraneous (ie. "left ACL injury" and "right ACL injury" were classified as two separate categories). We have reduced the number of distinct locations of injuries to a set of 45 unique values. At this point, it has been recommended to us that we run more EDA and conduct survival analysis with various groupings and subsets of injury type in order to see which model explains the most variation in our data.

3.4 Preliminary EDA

Most of our exploratory data analysis to this point has been with regards to the injury type and tendencies with regards to number of games missed measured by multiple different categorizations of time (games, dates on the Gregorian calendar, etc.)

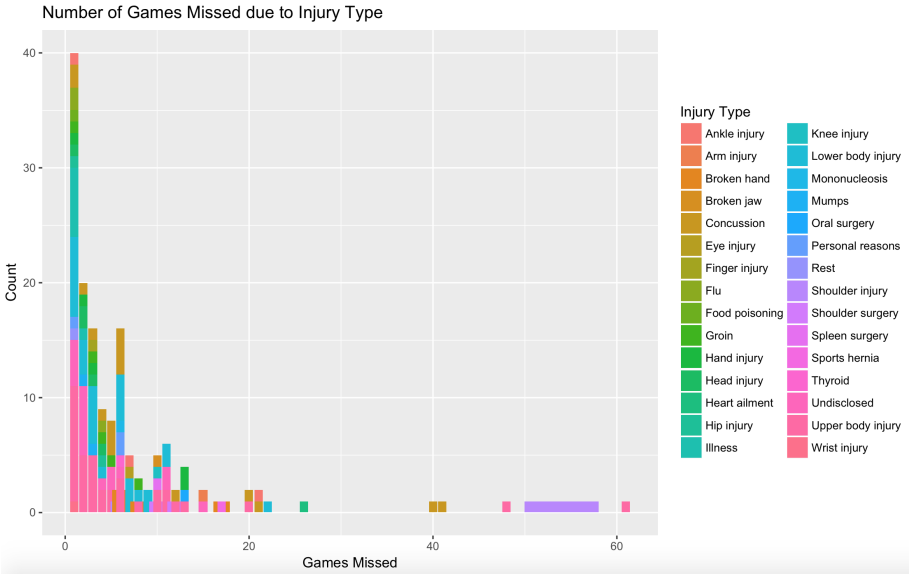

Figure 1: Games Missed Due to Injury Type

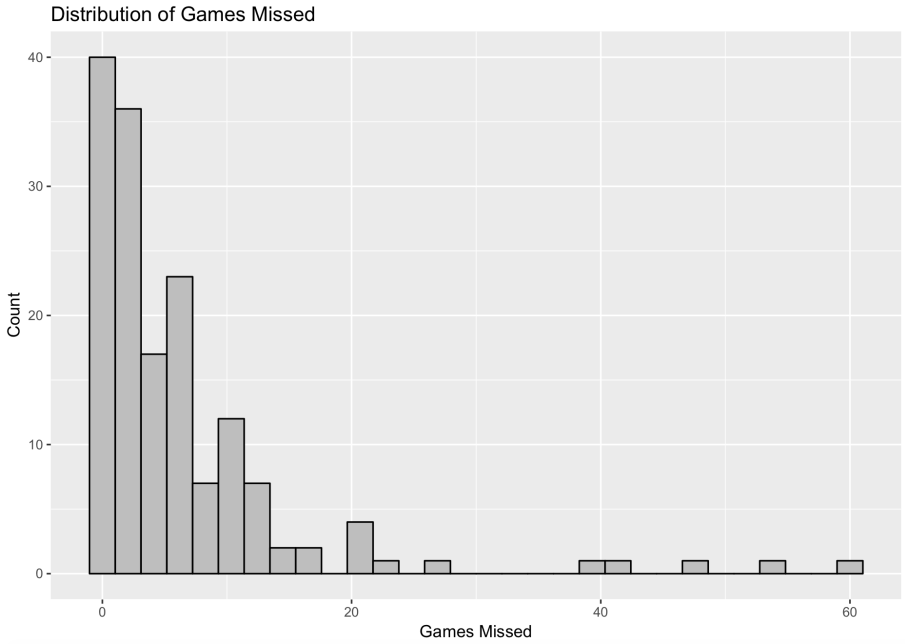

Figure 2: Distribution of Games Missed

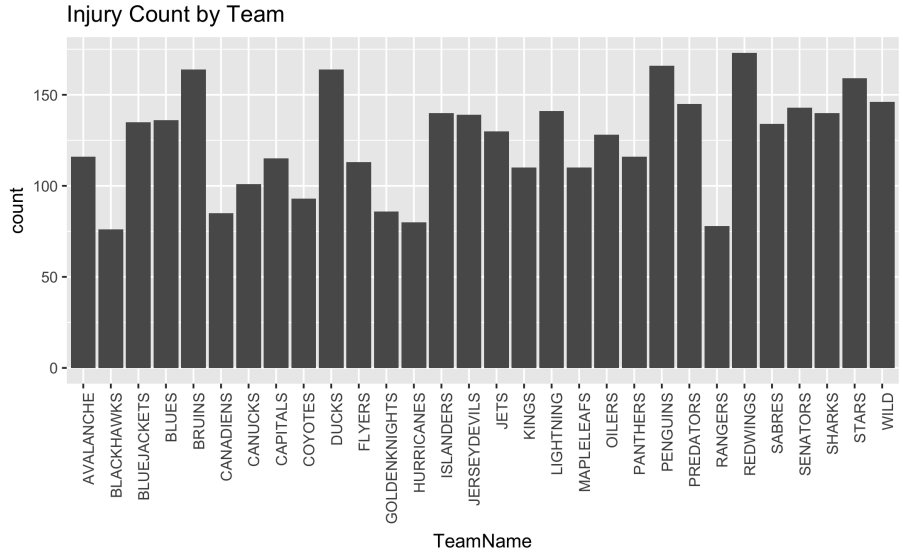

Figure 3: Distribution of Injury Count

4 Methods UPDATED

- **Survival Analysis** will allow us to model the expected duration of time for a player to come back from injury. Additionally, we can incorporate "treatments" such as injury type, team, previous injury recurrence, and later on demographic covariates. We will then test multiple injury classifications along the way. For example, we would test upper body injuries vs. all other classifications to see if there is any relevant information we can obtain from these models. There may not be one specific final model that we will use; it may be a comparison across various models that give us different information.

5 Results

We hope to create a plot (or multiple plots) that will demonstrate the proportion of players still injured based on an injury type (broad) or a specific injury.

To add: ** complications and/or missing data, EDA, modeling results

6 Discussion

main findings, strengths and limitations, future work

7 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