The Probability and Severity of Man Games Lost Due to Injury in an NHL Season

Jeremy Sylvain & Michael Schuckers

Jfsylv14@stlawu.edu

Abstract:

The goal of this project is to estimate the probability of an injury to an NHL player, and with that information try to predict the severity of the injury. To do this we model the probability that players will be injured by their ice time and position. This allows us to assess the distribution of injuries across a typical NHL team. Further, we model the expected amount of games lost when a player is injured. Our data covers multiple years of injury and time on ice data from mangameslost.com and hockey-reference.com. Using predictions from our model, we can predict the impact of staying healthy on a typical NHL team.

For our analysis, we collected several variables potentially useful in predicting injuries for regular season games. These variables included; games played during the regular season, time on ice per game, games lost due to injury, hits a player recorded during the season, and the number of shots blocked in the season. The data was collected for several seasons to gather information about injury rates through the season. To analyze injury likelihood, we built several regression models using the above variables. Separate models were built for forwards and for defensemen. Our approach is a two-stage one, we first modeled the probability that a given player would be injured in a given season. Next we model the length of injury given that an injury occurred. From these models, we can gain an understanding of the impact of these factors on injuries in the NHL.

The main outcomes from this project are that we built models of injury risk in the NHL and from those we can identify factors that impact the probability and longevity of an injury.

Introduction:

Among the many decisions that coaches face, lineup adjustments due to injury is one of the most common. Injuries can not only plague an individual player; it can affect the outcome of a whole team's season if that player is integral in the team's success. Traditionally the teams who have the most man games lost due to injury (MGL) are also the same teams that struggle to make or fail to make the playoffs each year. The correlation between injuries and a team's success throughout the season comes as no surprise as teams who lose players will continually call up players from their farm teams.

Seeing the correlation between injuries and success of a team, the players who are more resilient to injuries can be seen to have a higher value added to a team because they are less susceptible to injuries and will miss less time. Therefore, it is important to understand and model the probability of injury based on key factors of the game. Such factors include hits, blocked shots, and time on ice. The key is to use such variables to create a model that will indicate the level of these factors that will increase or decrease a player's probability of injury. Below we fit such a

model to data from recent NHL seasons using data from hockey-reference.com and mangameslost.com This is done by position for both forwards and defensemen.

Having built a model for injury probability, we create a model based on the same key factors to predict the duration of the injury. The goal is that the model will be able to approximate the length of an injury given the level of hits, blocked shots, and time on ice an individual player faces. Here we can also estimate the value to the team the player has based on their contribution to the team.

By using this two-step process we can hope to calculate the expected man games lost by multiplying the predicted probability of injury to the predicted length of injury. We find that the primary driving factor for the chance a player is injured is the amount of time per game that a player is on the ice, while injury severity is driven by multiple factors including on-ice actions including hits per game. These results are same for Forwards and Defensemen.

Data:

In order to fit models to injury data, we first obtained data on injuries in the NHL as well as some possible metrics that might impact potential for injury such as the number of hits a player carries out. To allow us to create stable estimates, we used data from the seven NHL seasons between 2009 and 2016. As mentioned above, the data that we collected came from mangameslost.com and hockey-reference.com. Below we list the metrics the we focused on as being possible impacts for the likelihood of a player being injured.

Table 1: List of variables and their abbreviations

Variable	Definition
INJ	Games a player lost due to injury
GP	Games played by a player
TOI-GM	Average time on ice played per game in all situations
HPG	Hits a player gives divided by games played
BPG	Shots a player blocks divided by games played
AGE	Age of player as if January 1st, of each season in years
AGE^2	Age of a player squared in years

The data was further broken down by forwards and defensemen (goalies are excluded from study). The data includes regular season games, as well as all situations during a game, that is, there is no difference between time on ice during even strength or shorthanded and power play time on ice. For this analysis, we excluded players whose total games in a season plus their games lost due to injury was less than ten games in a season. We felt that those players may not have been exposed to the rigors of the NHL for that season. Further we removed any player whose games lost due to injury was 82, as we felt that they did not play in game during the season therefore the injury did not occur in the current season we also did the same thing in the

2012-13 season with games lost due to injury that were 42 as this was the length of the shortened season. In total we had 6,630 observations and Table 2 has the breakdown of players per season.

Table 2: Numerical Summaries of Injury Data by Season

Season	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16
Number of Players	862	982	985	902	957	979	963
Percent of player with INJ>10	18.7%	21.9%	21.5%	13.6%	22.2%	22.1%	21.4%
Average Games Missed for Injured Players	5.86	7.05	7.25	4.01	6.83	6.67	6.65

Figure1: Histogram of INJ

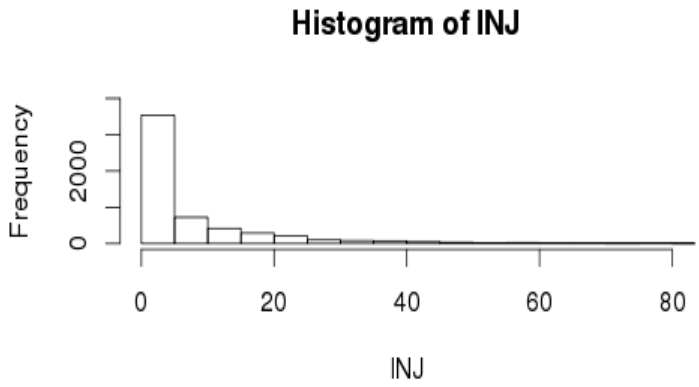
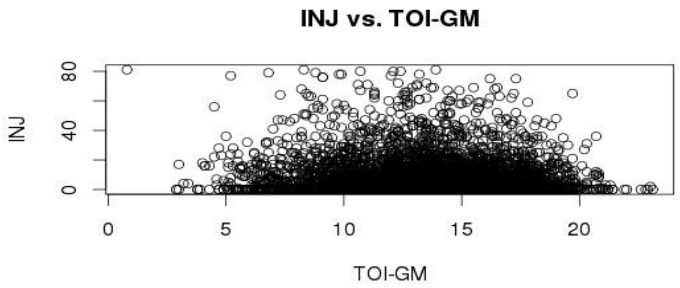


Figure2: Scatter plot of INJ vs. TOI-GM



Analysis

In this section, we introduce our approaches for modeling NHL injury data. For the probability of an injury we will use a logistic regression, and for length of the injury we will use a log-linear regression model. We include hits per game (HPG) and blocks per games (BPG) as covariates in each model as these involve collisions that can cause injury. Additionally, we include time on ice per game (TOI-GM) in our models to account for increased opportunities for injury. To account for increased susceptibility due to aging we include both age in years (AGE) and a quadratic term (AGE^2).

Injury Probability Model:

The first model in our two-model process is the injury probability model. The model, predicts the probability of injury based on the variables from the data. To do this we created to an indicator variable (INJPROB=1 if injured, 0 otherwise). Using the indicator as a response, we created a logistic regression model predicting the probability of injury based HPG, BPG, TOI-GM, AGE, and GP.

$$\mathbf{Logit(INJPROB)\sim HPG+BPG+TOI\text{-}GM+AGE+AGE^2\ (1)}$$

The model given in equation (1) was run for both forwards and defensemen for each season in the study, and then again run by position for all seasons combined. The model is set up in such a way that each variable should have a deleterious effect on the probability of injury. That is, we anticipate each predictor to have a positive coefficient so as the variable increases so does the likelihood of being injured throughout a regular season.

Injury Severity Model**:**

The second of the two models is used to model the severity of an injury given that a player was injured based on the same variables from the injury probability model. The model is based off of a Poisson distribution, where the response variable (INJ) takes on a logarithmic function. We do this so we can model the severity of the injury given that the player was injured during the

season. We use the Poisson model given in Equation (2) because of the increase in variation of

Forwards	TOI-GM	HPG	BPG	AGE	AGE^2
09-10	**0.1632	*0.3137	-0.0113	0.3002	-0.0037
10-11	**0.1300	-0.1498	-0.4418	-0.0124	0.0017
11-12	**0.1255	-0.0333	-0.0980	0.2574	-0.0034
12-13	**0.0919	0.0980	0.0040	*0.4798	*-0.0070
13-14	**0.1783	0.1737	0.2649	0.2678	-0.0031
14-15	**0.2137	**0.2580	0.3381	0.1840	-0.0014
15-16	**0.1826	-0.0584	0.0308	0.2026	-0.0026
All Seasons	**0.1489	0.0767	0.0477	**0.2358	*-0.0028
Defense	TOI-GM	HPG	BPG	AGE	AGE^2
09-10	*0.1129	*0.3753	0.1259	0.4488	-0.0069
10-11	**0.1411	0.1660	-0.1334	0.1034	-0.0006
11-12	**0.1503	*0.4184	0.1622	0.2756	-0.0029
12-13	0.0455	-0.0513	0.4614	0.3657	-0.0051
13-14	**0.1355	**0.4488	0.3376	-0.2547	0.0067
14-15	**0.1988	0.0543	0.1357	0.0002	0.0008
15-16	**0.1803	0.2565	-0.3823	*0.7199	*-0.0109
All Seasons	**0.1363	**0.2094	0.0647	*0.2627	-0.0030

injury severity as injury severity increases as is typical of count data.

$Log(INJ) \sim HPG+BPG+TOI\text{-}GM+AGE+AGE^2$ (2)

Like the injury probability model, we fit the model for each season of the data set separately, for forwards and for defensemen. Then we again fit the model with all seasons still differentiated by position.

The two models are used together in such a way that gives us the probability of an injury predicted by certain levels for the data, then based on the same levels as the probability equation we can predict the severity of the injury given that the player was injured.

Results:

Table 3: Coefficients and their significance for the injury probability model

NOTE: "*" level of significance, "**" ($p<0.01$) and "*" ($p<0.05$)

Table 4: Coefficients and their significance for the injury severity model

Forwards	TOI-GM	HPG	BPG	AGE	AGE^2
09-10	**0.0733	**0.1292	**0.3738	**0.3398	**-0.0050
10-11	**0.0505	**-0.1830	**-0.5150	**0.1950	**-0.0022
11-12	**0.0599	**-0.00489	**-0.6280	**0.4974	**-0.0079
12-13	**0.0361	**0.0939	**-0.6027	**0.5072	**-0.0073
13-14	**0.0438	0.0242	**0.5464	**0.2845	**-0.0036

Commented [js1]:

14-15	**0.0671	**0.2467	-0.0513	**0.2611	**-0.0034
15-16	-0.0048	**-0.0750	0.0084	**0.1722	**-0.0016
All Seasons	**0.0466	**0.0325	**-0.1300	**0.2894	**-0.0039
Defense	TOI-GM	HPG	BPG	AGE	AGE^2
09-10	**0.0872	**0.0955	**-0.3462	**0.7850	**-0.0124
10-11	**0.0436	-0.0188	-0.0255	**0.3341	**-0.0044
11-12	**0.0303	**-0.2320	**0.1862	**0.4876	**-0.0073
12-13	-0.0005	**0.1403	-0.1102	**0.4976	**-0.0082
13-14	-0.0033	**0.0906	**0.2256	**0.2533	**-0.0028
14-15	*0.0155	*0.0518	*0.1205	*-0.0930	**0.0026
15-16	**-0.0487	**0.2779	0.0237	**0.2630	**-0.0027
All Seasons	**0.0190	**0.0613	0.0071	**0.3180	**-0.0042

NOTE: "*" level of significance, "**" ($p<0.01$) and "*" ($p<0.05$)

We fit the Injury Probability and Injury Severity models given above for each year and for Forwards and Defensemen separately. The coefficients for our Injury Probability and Injury Severity Models by year can be found in Table 3 and Table 4.

Starting with the Injury Probability model for Forwards, we find that the most significant variable is TOI-GM. It was significant but small for all years. The remaining variables HPG, BPG, AGE, and AGE2 were not consistently significant for Forwards in this model. Blocks per game, were not significant for any of these seasons while HPG and Age were significant in two seasons and one season, respectively. A similar picture emerges for Defensemen, the coefficient for TOI-GM is significant in all but one season (2012-2013) while BPG is not significant in any of these models and HPG and Age are significant in at most two of the models.

Looking at the injury severity model (Table 4), the most significant factor in predicting the duration of an injury is split between AGE and AGE2. Both variables are significant in all models. The AGE variable has a positive coefficient meaning that the older a player gets the larger the severity of their injury (in terms of games lost). AGE2 is a very interesting variable not only is it statistically significant across all models, it also consistently has a negative coefficient. The negative coefficient implies that the more severe injuries happen to players that are in in the middle of their careers.

HPG is more frequently significant in the Injury Severity model, however so of the coefficients are negative implying that more hits lead to a smaller severity of injury. A similar occurrence happens with BPG as well where the variable is more frequently significant, but also some models have negative coefficients (Table 4).

Looking further at the injury probability models for all of the seasons within our study, it is clear to see that the variables chosen seem to have little predicting the probability, one of the ways we looked at strengthening the model was by limiting the number of games to be in the subset of data. The intention was to have players that have played a good portion of a season that had higher levels for each variable to strengthen the model. However, when we ran the model with a subset to include players who had played a minimum of 20 games, the significance of the variables that were once significant were no longer statistically significant at any level.

Conclusion and Discussion:

Looking at Table 3, the most significant variable is TOI-GM. It was almost always a statistically significant variable when predicting the probability of a forward or defenseman being injured. This is due to players being exposed to risk more as you increase your time on ice, which could create a potential problem for coaches and general managers as team's push for the playoffs. They need their star players on the ice however as they increase their time on ice they are also increasing their player's probability of injury.

A surprising variable that was always insignificant was BPG, between both positions BPG was always not statistically significant. There are many reasons that this could be caused by, one could be that blocked shots are defined by an opposition player getting in the way of a shot to prevent a shot from reaching the goal. The intentionality may imply that the blocker is using the body parts that are well padded thus mitigating the probability of being injured. A similar problem arises with HPG as well; a hit is recorded when one player initiates contact with their opponent. When a player is making contact with an opponent he is ready for the hit and is bracing for the hit, the ability to prepare for the hit allows for a player to better protect him from injury. It is rare in a hockey to see a person giving a hit out to be injured. It is much more likely for a player to be hit to get injured, therefore a stat like number of times injured in a game would be a better predictor of injury probability

If you compare the significance of the coefficients between positions in both models you can see that the forwards subset of data has more consistently significant p-values in their models especially when looking that the injury severity models (Table 4). One of the main reasons for that is the sample size, the forwards subset sample size is 4,344 players whereas the defensemen subset is only 2,286 players. The sample size difference could be the reason for the difference in the p-value's significance. However, the problem of the different sample sizes arises naturally in the NHL, on a 20-man roster only about 6 players are defensemen, and the rest are forwards, therefore there are disproportionately larger number of forwards to collect data from than there are defensemen.

The goal of this was to create a framework for predicting the probability of an injury occurring, and also the severity of that injury. The study introduced a two-part model that A. tries to model the probability of an injury occurring, and B. tries to predict the severity of the injury in terms of man games lost. While the results are not quite as robust as one would, there is a lot of stuff to be optimistic about. One is that we have laid the ground work for modeling injury frequency within the NHL. Another is that with this framework we have plenty of room to build future models and continue to model the behavior of injuries in the NHL, there are several extensions we would like to build off of using our model as the foundation.

One of the most important extensions is the situation of the game, this includes; even strength, power play, shorthanded (both five on four and five on 3), and also overtime. The current models only work for even strength situations, and with the addition of those key parts of the game it would be very interesting to see how the probability of being injured would change and also how the severity of the injury would change.

Another extension would be to add to the variable that are already in the model. One that might be beneficial would be a player's height and weight, and also maybe an interaction with age. It would be very interesting to see how that would affect the probability of injuries. One would expect to see the players have a higher probability of injury if they are smaller.

Along those lines would be, to see if the player has been injured before, and in addition to that it would be interesting to weight that historical data as a distance from the most recent season in the data set. The extensions listed above would allow us to gain a better insight into the frequency and severity of injuries during the NHL regular season.

Acknowledgement: We would like to acknowledge both www.hockey-reference.com as well as www.mangameslost.com for their excellent data that we were able to use to conduct this study.