How Tough are Hockey Players?

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

Jeremy Sylvain & Michael E. Schuckers

St. Lawrence University

Abstract:

The goal of this project is to estimate the probability of an injury to an NHL player, and with that information predict the severity of the injury. To do this we model the probability that players will be injured by their ice time and position and we use on-ice events as part of these models. 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 NHL coaches and general managers face, lineup adjustments due to injury is one of the most common. Injuries can not only plague an individual player; they can affect the outcome of a whole team's season. 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. For example, for the 2016-17 NHL season the correlation between the percent cap hit of injured players per game and team points per game was -0.31[1]. Thus,[1] it is clear that injuries have an impact on team performance.

[1] Statistic from http://nhlinjuryviz.blogspot.com/2016/10/201617-team-injury-breakdowns.html

Seeing the correlation between injuries and success of a team, it is important to understand and to model the probability of injury based on key factors of the game. The factors we will use include hits, blocked shots, and time on ice. The key is to use such variables to create a model that will indicate the importance of these factors and how they 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 prediction 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 and validate our results, 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 that we focused on as being possible factors in 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

We divided our data by position into forwards and defensemen (goalies are excluded from study) since we felt that forwards and defensemen were affected differently by the predictors in the models. These data include regular season games, as well as all situations during a game, that is,

we did not distinguish between time on ice during even strength or shorthanded and power play time on ice. 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 48 as this was the length of the shortened season. We did not remove the 2012-2013 season because we were looking at time on ice and not games played so we saw the minutes played during a game were a valuable addition to our analysis regardless of how many games were played in the season. In total, we had 5,664 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 Players with INJ≥1	43.2%	49.1%	52.9%	40.2%	52.9%	51.5%	51.0%
Percent of Players with INJ≥5	31.6%	33.7%	38.0%	24.7%	38.1%	34.7%	34.5%
Percent of Players 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
Median Games Missed for Injured Players	0	0	1	0	1	1	1

From the summaries in Table 2, we can see that approximately 20% of NHL players are injured more than ten games in a season. Similarly the average games missed for each player was in the neighboorhood of six or seven games missed, the median games missed was between zero and one. The larger magnitude of the mean number of games missed relative to the median number of games missed suggests that there is a long right tail on the distribution of games missed per player. We see exactly that in Figure 1, the histogram of games missed per player per season. Noteworthy among these results by season is the results for the 2012-13 shortened lockout season. During that campaign, the percent of significant injuries, those lasting more than 10 games, was fewer than the average games missed and was subsatintially less.

Figure2 is a scatterplot of of the relatioship between a player's average time on ice per game and their number of games missed in that season.

Figure 1: Histogram of INJ

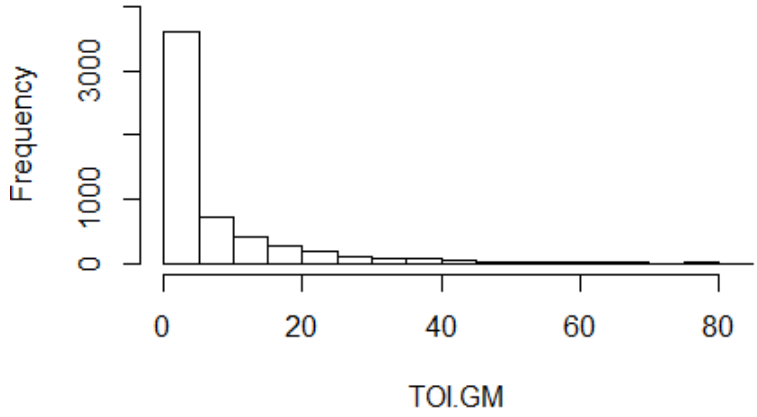
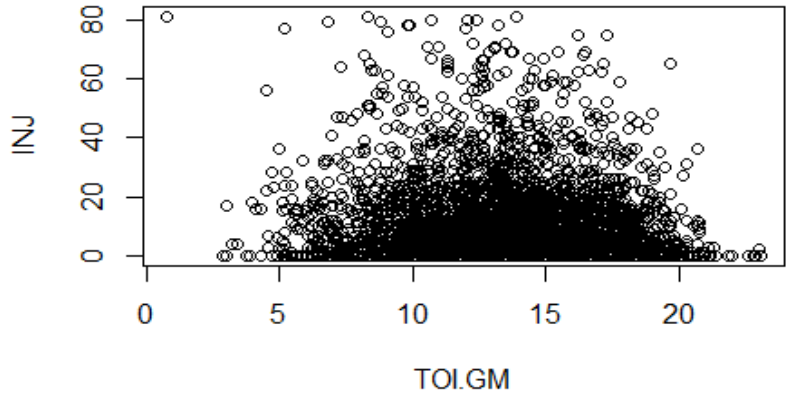


Figure2: Scatter plot of INJ vs. TOI-GM



Analysis

In this section, we introduce our approaches for modeling NHL injury data. This will be done in two parts. For the probability of an injury we use a logistic regression, and for length of the injury we 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 may 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 injury susceptibility due to aging we include both age in years (AGE) and a quadratic term age in years squared (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.

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

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.

Increased Games in the Probability Model:

When a team starts to get into a comfortable spot towards the end of the season, coaches might make the decision to rest the players who might see a lot of ice time who might be of verge of a long-term injury, or players who are sick and cannot play at their full potential. Such incidents get rolled into a game lost due to injury. Therefore, a player who might not have actually been injured during the course of actual game played might record a game lost due to injury. To try to capture this possibility, we increased the response variable to the probability of missing at least 5 games (INJ≥5) we call this INJPROB5. We used the same factors that are found in equation (1) for the model for this addition response. The results for this model can be viewed in Table 5.

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 on a Poisson distribution, where the response variable (INJ) is modeled using a logarithmic link 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 injury severity as injury severity increases as is typical of count data.

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

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

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, Table 4 and Table 5.

Starting with the Injury Probability model for Forwards, we find that the most significant variable is TOI-GM. It was significant 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. Thus, for both forwards and defensemen, the number of blocked shots per game and the number of hits per game are not significant predictors of a player missing more than one game.

Table 3: Coefficients and their Significance for the Injury Probability Model (INJ≥1)

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

Looking at the injury severity model, estimates found in 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 some of the coefficients are negative (and significant) 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). As was the case in the injury probability model, average time on ice per game is a consistently significant predictor of injury severity.

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
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

Looking at the estimates in the probability model for predicting missing more than 5 games, Table 5, we see in the table TOI.GM is still the most significant predictor in the model predicting the probability of missing 5 or more games. It was significant for all season except for the 2012-2013 for forwards and only significant for 4 seasons for defense. HPG became insignificant for the forwards in the model, and the defense only had two seasons where HPG were significant at predicting the probability missing more than 5 games during a regular season. Like the probability of missing at least one game, BPG was not significant at predicting the probability of an injury during any season, making it an ineffective at predicting the probability of the probability of missing 5 or more games due to injury. AGE and AGE^2 were both only significant in one season each at predicting the probability of missing 5 or more games due to injury, which is slightly worse than the probability model for predicting at least one game lost due to injury where AGE and AGE^2 were significant in 2 seasons each.

Table 5: Coefficients and Their Significance for the Injury Probability Model (INJ≥5)

Forwards	TOI.GM	HPG	BPG	AGE	AGE^2
09-10	**0.100	0.244	-0.090	0.389	-0.005
10-11	**0.103	-0.129	0.007	0.008	0.004
11-12	**0.103	0.086	-0.109	0.234	-0.003
12-13	0.053	0.161	-0.440	*0.565	*0.009
13-14	**0.078	0.127	0.452	0.155	-0.001
14-15	**0.110	0.177	0.276	0.297	-0.004
15-16	**0.097	-0.106	0.208	0.317	-0.004
All Seasons	**0.090	0.068	0.010	0.261	-0.003
Defense	TOI.GM	HPG	BPG	AGE	AGE^2
09-10	0.088	0.216	0.060	0.414	-0.007
10-11	*0.088	0.229	-0.146	0.226	-0.002
11-12	0.035	0.070	0.374	0.178	-0.002
12-13	0.012	0.0149	-0.332	0.607	-0.010
13-14	0.041	*0.330	0.408	-0.195	0.005
14-15	*0.120	0.199	0.027	0.015	-0.001
15-16	*0.124	0.288	-0.079	*0.791	*-0.012
All Seasons	**0.073	**0.205	0.014	**0.280	-0.003

Application.

To apply the models to a real-life situation, we decided to use a 23-man roster from the Ottawa Senators during the most recent season (2016-2017). We took the probability models encompassing both forwards and defensemen and fit these models to the Senators data. To gain an understanding of the variation in our models, we simulated the injury impact of 1000 2016-17 seasons for the senators. This allows us to create a stable distribution of predicted man games lost due to injury. The results are shown below in Figure 3(Probability of missing at least one game), and Figure 4 (Probability of missing at least 5 games). The actual man games lost due to injury used in the simulation was 133, which is represented by the vertical red line, Ottawa's actual man games lost during the 2016-2017 regular season was 211 however, 78 of the 211 came from Clarke MacArthur who was not injured during the regular season and therefore was not considered for this simulation. For our model, we used we had 2 different simulations one for the probability of missing at least one game, and the probability of missing 5 or more games. For the probability of missing at least one game of the regular season 133 fell within a 95% confidence interval between 70 and 156 games lost due to injury. When you increase the probability model so that it is predicting the probability of missing 5 or more games (INJ≥5) the model loses some its predicting power. The 95% confidence interval for the simulation was between 44 and 126 and where the Ottawa Senators only lost 133 games due to injury our model predicted lower than the actual.

Figure 3: Distribution of Predicted Man Games Lost for the Ottawa Senators 2016-17 Season (INJ≥1)

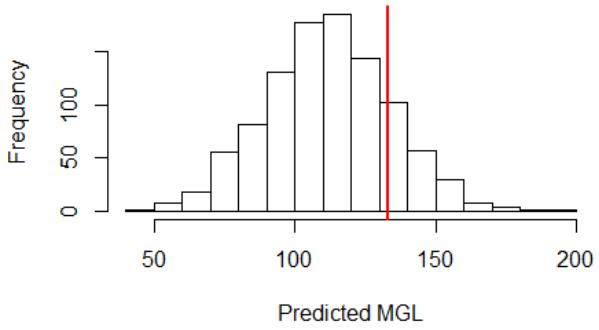
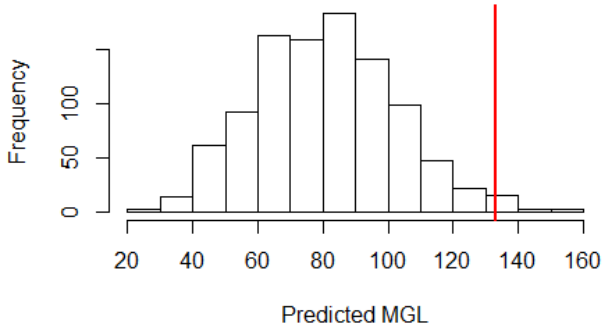


Figure 4: Distribution of Predicted Man Games Lost for the Ottawa Senators 2016-17 Season (INJ≥5)



You can see this by looking at the histograms, for both models. You can see in the probability of missing one game (top) that the actual number of missing games is much closer to the center of the distribution, where the probability of missing at least 5 games(bottom) was much farther right and at the end of the distribution. This can imply that the factors in our model are better at predicting probability of lower end injuries (games lost < 5). Where the model for predicting lengthier injuries could be much better explained by different factors.

Conclusion and Discussion:

In this paper, we have built statistical prediction models for injuries during a single NHL season. We fit logistic regression model for the probability of missing at least one game and logistic models for the probability of missing 5 or more games. To consider injury severity we fit log-linear regression models to the number of games lost due to injury by a player in a given season. Separate models for forwards and defensemen. For all of these models, the most consistently significant variable is TOI-GM. It was almost always a statistically significant variable when

predicting the probability of a forward or defenseman being injured by year and it is always significant when we modelled the data in all seasons. This may be due to players being exposed to more risk more as their time on ice increases, 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 blocks per game, BPG. Between both position types, BPG was never statistically significant in predicting whether or not a player would be injured. There are many possible explanations for this including the possibility that those who often block shots are proficient at that task and thus, less susceptible to injury from blocking shots. 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 the number of hits received might be a better predictor of injury probability. Another useful metric for modelling injuries might be penalty minutes which could be a proxy for reckless or borderline behavior by players.

If you compare the significance of the coefficients between positions in both models you can see that the forwards have more consistently significant p-values in their models especially when looking that the injury severity models (Table 4). One possible reason is the sample size, the forwards subset sample size is 3703 players whereas the defensemen subset is only 1961 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 7 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. We introduced a two-part approach that modelled the probability of injury and severity of the injury in terms of man games lost. While the results are not quite as robust as one would like, we have laid the ground work for modeling injuries in the NHL. The framework here allows for future models with additional predictors among these additional predictors we would like to investigate in the future are time on ice at difference strengths (even strength, power play, shorthanded) as well as penalty minutes per game and hits received per game. The current models only uses all strengths and the addition of more granular strength data could be vital to improving the how we model the probability of injury and the severity of injury. We conjecture that additional penalty kill time leads to increased probability of injury because of the exposure to shots and the increased shot-blocking responsibilities of penalty killers. Other metrics about the player being considered such as height and weight might beneficial. We might expect smaller players are more susceptible to injuries but a full investigation of that would be necessary. Addressing how these variables are differentially impact our responses at different ages is another avenue we hope to pursue.

Along those lines, we would like to examine the impact of previous injury on a player's susceptibility to future injuries. This and the other extensions 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. Without the contributions of these two sites we would have been unable to complete this study. Nathan Currier of mangameslost.com, in particular, was especially generous with his comments and data for this project.