How Does Player Injury Impact Game Success in the NFL?

Matthew Guo, Louis Wong, Kerri O'Brien, Eric Crouse

Problem and Background

Sports injury can be an unpredictable factor in player and team performance. This paper attempts to make sense of injury data to evaluate its relationship with team performance and valuation. This investigation's critical question is: How is NFL player injury associated with team performance and valuation? Various machine-learning models were used to examine these relationships, including multiple logistic and linear regression.

While injuries in the National Football League (NFL) have been quantified in the past, their use in predictive models has largely remained inconclusive (Clubb, 2021; Langaroudi & Yamaghani, 2019). Our goal was to summarize the impact of injuries in NFL football on team success and to transform the effects of various injuries into actionable insights for NFL teams and fantasy football managers. We analyzed various types of injuries to players of different positions. We sought to predict return-to-play time and effects on team offense/defense effectiveness and wins/losses based on data available at the time of injury.

Our findings impact team managers, fantasy football managers, and NFL fans. The more sophisticated the model, the more efficient teams can be in resource allocation, and the better the audience can be informed regarding sports tourism and the betting industry. Put simply, we define a data-driven approach to understanding what it means when an NFL player becomes injured.

Datasets

We had seven raw datasets regarding NFL statistics from 2020 to 2022 inclusive. This time range includes the sports industry's recovery from COVID-19, so this analysis should be generalizable to present and future games. The data were mostly time series by week/season, but each dataset had specific granularity and dimensions. Most datasets were split by player, team, week, and year, averaging about 15,000 rows per file.

The data came from The Pro Sports Transactions Archive, Kaggle, and various NFL databases. The raw injury dataset was ingested as a time series of players moving from “injured” to “not injured” and vice versa. We transformed this data by tracking each player over week/season and adding an “injured” boolean to denote whether they were injured for a given week. In this format, we could easily discern whether a player was injured during a specific date – not just when they were moved in/out of play. There were 15,611 observations from 2020 to 2022 inclusive.

Two of the datasets tracked players over week and season. Attributes include metrics on passes, rushing, points, etc. The datasets even had the field surface material, temperature, humidity, and wind speed. Based on the game date, we added season, year, and week numbers as a composite key to link to the other datasets – especially the injury data. In addition, we had a static cross-section of NFL players' heights, weights, birthdates, universities, and positions.

Methods

The injury and success data were cleaned, transformed, and linked to a unique list of NFL players. Once the data had been extracted, loaded, and cleaned, the analysis employed machine learning models to determine the effect of injury on various success metrics. The models were identified by iterating through them in the scikit-learn Python library and evaluating their accuracy and precision. Each model had its success metric (e.g., root mean squared error for regression and Receiver Operating Characteristic Area Under the Curve for binary classification).

Discussion

Injuries' Effects on Team Performance

First, there was the open-ended question of how injuries affect team performance. The answer would require some data manipulation and exploratory analysis. After target data was selected based on relevant testable attributes, the data was cleaned to impute missing values with the mean and remove duplicate records. To understand the impact of injuries on team performance, each team's "injury status" would need to be identified for each game.

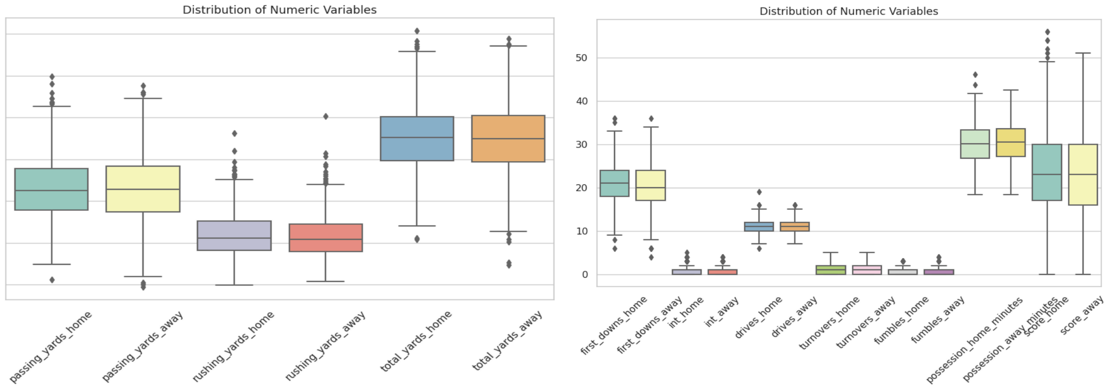
We determined that the most effective way to get this information was to iteratively update each record in the 'games' dataset (containing one row for each game played) to include T/F values for each position indicating their injury status (each cell would read True if the specified team had an injury at that position, and False otherwise). This required merging two other datasets to identify each player's position and team. Then, depending on whether the home or away team had injured players before a matchup, the attributes of the created 'injury' column were populated accordingly. This was done for both home and away teams. A boolean variable called 'home_win' was created, with a value of 1 if the home team won and 0 otherwise. Finally, possession time was converted to datetime.time datatype (periodically temporarily converted to integer values in minutes to obtain the average).

Once all data was clean, properly formatted, of the correct data type, and valuable attributes were created, the data was ready for analysis. Binning was used to establish three groups/classifications of records: those where the home team suffered no injuries ("Healthy"), those where the home team suffered exactly one injury at any position ("1 Injury"), and those where the home team suffered two of more injuries ("2+ Injuries"). Additionally, data was grouped by specific injured position (all records where the home team had an injured quarterback (QB), running back (RB), etc.). Together, these made up what will henceforth be called "injury statuses."

We created box plots to get an idea of the distribution of each numerical variable. The box plots did show some unusual values and outliers (such as -1 rushing yards for some games, etc.); however, cross-referencing revealed that reputable sources like ESPN corroborated the data, so these values were natural. Given the natural variability of sports performances, we determined that these values should be considered along with the rest for this study.

Figure 1

Distribution of Numeric Team Performance Metrics

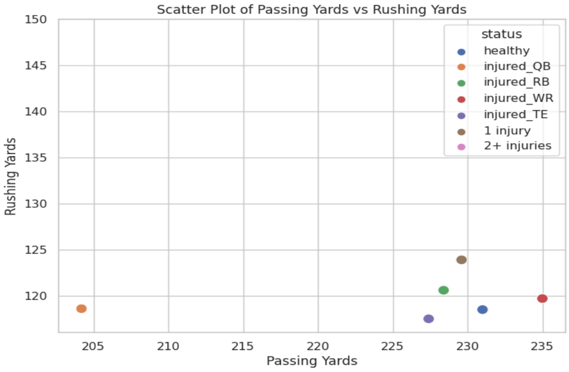


We calculated mean values for each numeric attribute and compared them across injury-status categories. This revealed nothing statistically significant among categories with sufficiently large n (four injured positions were not considered viable due to n < 20). The most significant variation was between healthy teams' passing yards and teams with injured QBs (about 27 yards). It is worth noting that even when no statistically significant variation was present across all injury statuses, teams with injured quarterbacks consistently had the lowest performances.

The first statistical test performed was a point-biserial correlation (using Python's scipy.stats pointbiserialr) between dichotomous injury-status variables (home_QB_injured, home_WR_injured, etc.) and continuous numerical variables (yardage, first downs, interceptions, fumbles, turnovers, etc.). This test returned only two statistically significant correlations: those between injured quarterbacks and passing yards and between injured quarterbacks and first downs. Both were weak negative correlations, which is consistent with intuition.

Figure 2

Mean Passing and Rushing Yards for Each Injury Status



The right scatter plot was made with axes 'Passing Yards' and 'Rushing Yards,' and means were plotted to visualize the differences among categories. An injured QB is associated with a noticeable drop in average passing yards (about 27). However, no other category shows significant variation in mean passing or rushing yards, which is interesting.

An attempt was made to compare individual team-level statistics (how each team's performance under different injury conditions differed from their performance when healthy). However, it quickly became apparent that the sample size was too small to determine meaningful results.

The following variable explored was possession time, which had not previously been included in the numerical calculations (other than the initial univariate box plots). The resulting averages are below.

Figure 3

Possession time descriptive statistics

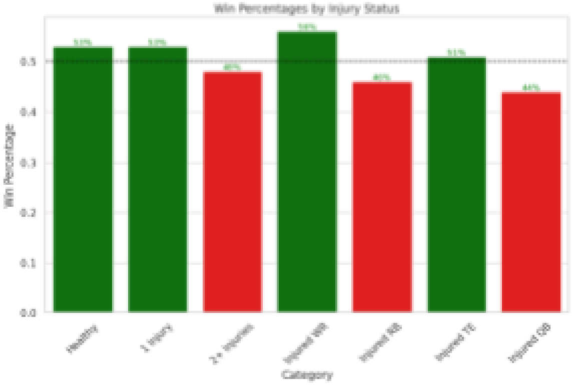
Max	Min	IQR	Healthy Mean	1 Injury Mean	2+ Injuries Mean	Injured QB Mean	Injured RB Mean	Injured TE Mean	Injured WR Mean
00:46:04	00:18:25	00:06:36	00:30:04	00:30:10	00:29:44	00:29:16	00:29:36	00:29:41	00:30:40
n=769	n=769	n=769	n=296	n=305	n=168	n=77	n=158	n=189	n=218

Given the IQR, none of these variations are significant. However, once again, the injured QB category has the worst performance of all groups, including games played with two or more injuries. To confirm this, an analysis of variance was performed using Python scipy.stats, which yielded an F-Statistic of 1.5344, P-Value of 0.1632, and a failure to reject the null hypothesis.

Win percentages were also calculated across all injury statuses. No significant variation was found. Nevertheless, teams with injured quarterbacks performed the worst,

Figure 4

Win Percentages for Each Injury Status



winning just 44% of home games (compared to healthy teams at 53% and teams with multiple injuries at 48%). While not statistically significant, weak evidence continues to support the notion of quarterback being the most impactful position.

A multiple logistic regression model was constructed with an 80-20 training-testing split using Python sklearn with inputs strictly limited to home and away injury statuses and dichotomous outcome 'home_win' (T/F). It resulted in a 58% accuracy, as shown in the following table.

Figure 5

Performance Metrics of Multiple Logistic Regression Model

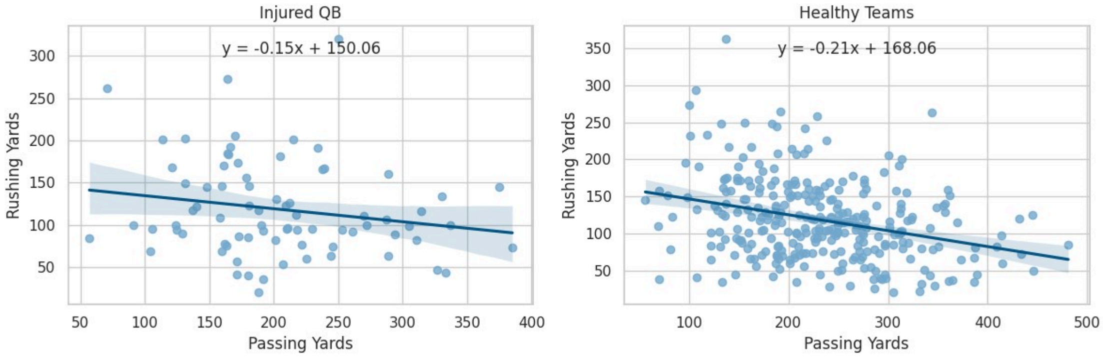
	Precision	**Recall**	**F1-Score**	**Support**
False	0.54	0.51	0.52	69
True	0.62	0.65	0.63	85
Accuracy			0.58	154
Macro Avg.	0.58	0.58	0.58	154
Weighted Avg.	0.58	0.58	0.58	154

Other models were constructed similarly, but they all achieved low accuracy. There are two reasons why this might be the case: first, the data we were able to source for this project needed to be sufficiently large. With records of 769 games, the number of records for each category decreases with the specificity of the category, and the sample size quickly becomes too small to be meaningful. The second reason is apparent: NFL matchup outcomes are impossible to predict perfectly, and even the best models from professional analytics teams that consider many variables can only achieve around 65% accuracy (www.nfeloapp.com).

With these attempts exhausted, the exploratory analysis of injuries' impact on overall team performance turned to the only statistically significant relationships it could find. Intuition guided a decision to plot passing yards vs rushing yards for a few different categories (Injured QB being among them) and see how the relationships differed.

Figure 6

Passing and Rushing Yards Trends



Injured QB

Healthy Teams

This plot was chosen for two reasons: firstly, rushing yards were demonstrated to be remarkably consistent among all categories–likely since they depend more on collective team performance than passing yards do–so they provided a baseline from which to examine passing yards. Secondly, the relationship between passing and rushing yards is well-known and easily understood (rookieroad.com).

The gentler negative slope on the Injured QB graph of Figure 6 represents a more significant variation in passing yards per unit change in rushing yards among teams with injured quarterbacks than among healthy teams. The greater range of passing yards among healthy teams indicates higher peak performance and higher average performance in terms of passing yards. This, combined with the significant negative correlation found between injured QBs and passing yards, offers some limited support for the notion that in addition to the weaker overall performance of backup quarterbacks, there is more variation in backup QB performance than in starting QB performance. This is also consistent with experiential knowledge and informal observations.

Predicting Individual Player Injury

With weekly game injury data (year, week, player, injured, injury type) merged with player data (position, date of birth, height, weight), the goal is to predict whether a player will get injured. Before importing the dataset into Python, it was preprocessed to merge the two datasets and handle missing values and outliers. The player dataset included date of birth, which was used with the game injury dataset to calculate their age. There were outliers in age that were reviewed individually and either verified or updated with the correct information. For missing values within age, height, and weight, the averages of each of those variables replaced the null value. The averages were calculated by position to create a more accurate dataset since the players' body sizes and ages are correlated to their position. While processing the data, the variables with string values were converted to numerical values in the dataset before being imported into Python. Positions were sorted alphabetically, and numbers 1 through 8 were assigned. Whether a player was injured for that week's game was already a bit field but was converted from True and False to 1 and 0, respectively.

Figure 7

Averages of Height, Weight, and Age by Position used in Missing Values

Position	Average Height (in)	Average Weight (lb)	Average Age
Kicker (K)	72.19	202.58	26.0
Quarterback (QB)	75.43	224.97	29.0
Running Back (RB)	70.73	214.48	26.0
Tight End (TE)	76.54	254.26	27.3
Wide Receiver (WR)	72.40	200.32	29.9

To predict whether a player will get injured, the statistical model chosen was multiple regression analysis, specifically logistic regression. Logistic regression was the best fit since the dependent variable (injured) is categorical (True/False). The independent variables used in the analysis are season years (2020-2022), regular season weeks (1-18), position, age, height in inches, and weight in lbs. Initially, the use of injury type wanted to be included but skewed the prediction since the data only existed for injured players, which was the outcome variable. For the logistic regression, all variables used were numerical values (converted from string values where necessary) and not highly correlated with the outcome variable. The table below was generated to view any existing correlations between the position and the outcome variable.

Figure 8

Percent of injuries by position for the 2020-2022 regular seasons (weeks 1-18)

Position	Outcome		Row Totals
	Not Injured	Injured	
Kicker (K)	10%	90%	10
Quarterback (QB)	88.6%	11.4%	1,933
Running Back (RB)	84.9%	15.1%	3,946
Tight End (TE)	78.9%	21.1%	3,118
Wide Receiver (WR)	84.2%	15.8%	6,405

Grand Total	83.7%	16.3%	15,611

Since injury type cannot be used as an independent variable as it is the same as the outcome variable, just with more detail of injury type, the below table compares various types of injuries sustained by players, such as (e.g., knee, ankle, concussion) to position type (e.g., quarterback, running back, wide receiver).

From personal knowledge of football policies, players who sustain a concussion will have more significant impacts on the game metrics as the league has a concussion protocol that limits the player's playing time and can lead to them missing games. This protocol makes the outcome more consistent across positions as it applies to each player the same. The table below corroborates the assumption that injury types vary based on position type. No injury type appears to be overly consistent across the position types other than the injured reserve, which refers to a player who has a non-specified football-related injury and will need to miss a minimum of a few weeks. The movements each position requires are different and, therefore, will likely have differing impacts depending on the player's position and individual playing style.

Figure 9

Injury Types by Position

Injury Type	K	QB	RB	TE	WR	Grand Total
abdominal					0.2%	0.1%
ankle			16.7%	15.3%	9.0%	11.9%
back			0.2%	1.0%	1.6%	1.0%
calf		7.9%	1.3%	1.0%		1.1%
chest					0.1%	0.1%
concussion			5.4%	8.6%	10.8%	8.1%
foot			6.7%		9.3%	5.5%
groin			3.9%	5.2%	6.9%	5.2%
hamstring			12.8%		5.1%	5.2%
hand		9.3%	3.7%	4.6%	0.2%	2.9%
hip				0.8%	3.3%	1.6%
IR		38.6%	30.7%	24.2%	36.2%	31.8%
knee	100%	9.3%	7.6%	28.8%	3.7%	12.0%
neck		2.9%	1.7%	1.7%		1.1%
non-football injury/illness		4.3%	7.4%	5.7%	6.6%	6.3%
ribs		2.9%		0.2%	0.5%	0.5%
shoulder		25.0%			4.7%	3.8%
thigh			1.9%	3.1%	1.7%	2.0%
Grand Total	100%	100%	100%	100%	100%	100%

The data was merged with Python pandas, and logistic regression was applied using scikit-learn with a training-testing partition of 70:30. Since there were multiple independent variables and they were all numeric, it was necessary to standardize the variables. We used scikit-learn's StandardScaler to do z-score normalization (i.e., z = (X–mean(X))/std(X)). This process was necessary before running logistic regression so the data follows a standard normal distribution to prevent some features with larger values from dominating the training process (i.e., weight).

Running Classification Report by scikit-learn provided metrics to evaluate the model's performance. The model had a high test accuracy of 83.2% and a precision of 83%. Accuracy is a good indicator and informative of the positive performance of the model. While we could not include the injury type data, the model is still helpful in predicting if a player will get injured and uses a variety of independent variables (year, week, position, age, height, weight) to predict it.

Figure 10

Classification Metrics

	Precision	Recall	F1-Score	Support
Not Injured (0)	0.83	1.00	0.91	3899
Injured (1)	0.00	0.00	0.00	785
Accuracy			0.83	4684
Macro Avg	0.42	0.50	0.45	4684
Weighted Avg	0.69	0.83	0.76	4684

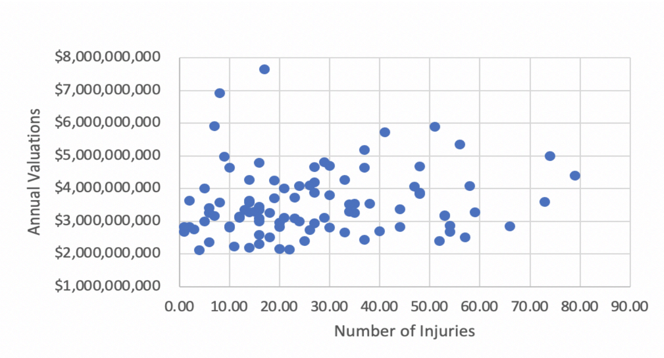
Team Valuation

A rigorous analytical process was undertaken to assess the correlation between injury data and the valuation of NFL teams. The data was meticulously prepared and transformed using Python's pandas and statistics libraries, facilitating the generation of comprehensive statistics across various categories of teams with differing injury statuses. This meticulous approach created a novel dataset encompassing detailed information about each game played over the past three years. This dataset included crucial details about the home and away teams, their respective injury statuses, and their corresponding performances. Introducing these new injury-related attributes allowed for a meticulous examination of how injuries influence the valuation of NFL teams. Preliminary findings have demonstrated noteworthy correlations between specific injuries and the overall valuation of teams. Notably, it was observed that certain key positions hold a substantial sway over a team's valuation, with quarterbacks (QB) emerging as a standout influencer. This initial discovery suggests that the presence or absence of injuries in critical positions can significantly impact the overall value of an NFL team.

Bivariate Analysis

Figure 11

Number of Team Injuries and Annual Valuations



Many mixtures of two variables were explored during this step to attempt to uncover predictive relationships. The main focus of this analysis was to answer the question: “Does an injury of a team impact the annual valuation of an NFL team?” To explore this question, data across 2020-2023 was analyzed to build the scatterplot, with the number of injuries as the independent variable and the annual valuation in dollars as the dependent variable. This visualization does not identify a strong relationship between the number of injuries and the annual valuation of an NFL team.

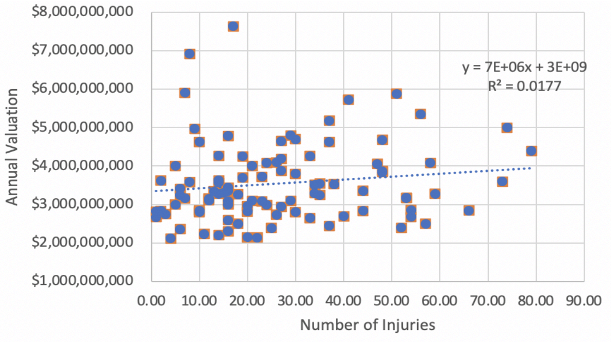
Bain & Company claims, "More than half of the deals made in sports over the past five years were investments in teams, leagues, or media rights.” Private Equity and Venture Capital Firms mainly made these investments. Bain mentioned that the shift to streaming, deeper fan engagement, higher value for advertisers, the rise of sports betting, and technology as a force multiplier are some essential opportunities in sports for potential investments to happen. However, this graph does not indicate that sports injury would heavily impact a team's valuation and overall potential investment. Due to this, a deeper dive was taken into metrics to quantify the strength of this relationship.

Linear Regression Model

Despite the weak relationship between the annual valuations of an NFL team and the number of injuries, this is a weak linear regression between the two variables. Python was used to clean and manipulate the data, and the actual visualization was finished in Excel. The model and visualization used Python packages pandas, pyplot, matplotlib, and scikit learn.

Figure 12

Linear Regression: Number of Team Injuries and Annual Valuations



The regression line had an R^2 score of 0.0177, which entails a very weak positive linear correlation. This could be explained by the fact that a game's injuries are unpredictable and not necessarily dependent on a player's skill set but rather just the nature of a contact sport like football. There is some argument to be had that since it is a very weak positive correlation, an increase in injuries correlates with an increase in valuation. Usually, the games with big-name teams have more injuries, which attract more fan engagement (a driver for a sports team's valuation). From an NFL team's perspective, it is settling to know that injuries do not have much impact on a potential investment; in the short run, it could help more fan engagement. In the long run, there is an argument that too many injuries can decrease fan and investor confidence.

Conclusion

This study delves into the intricate dynamics between NFL player injuries, team performance, and valuation through machine learning models, including linear and logistic regression. Some significant correlations were found, including the adverse effects of injured quarterbacks. We also successfully predicted individual player injuries with an accuracy rate of 83.2% based on factors such as position, age, height, and weight.

Our findings underscore the importance of data-driven approaches in understanding the intricate relationships that shape the NFL landscape. These outcomes offer valuable insights to team managers, fantasy football enthusiasts, and stakeholders in the industry. As the NFL continues to evolve, this research serves as a stepping stone for further exploration into the complexities of injuries in professional football and their far-reaching implications.

References

Clubb, J. (2021, August 23). *Does injury availability affect your team's chance of success?* Global Perf Insights. https://www.globalperformanceinsights.com/post/does-injury-availability-affect-your-team-s-chance-of-success#:~:text=Injuries%20not%20only%20make%20athletes,athlete%20and%20potentially%20in%20his

Greer, Robby. "NFL Model Performance." Games, www.nfeloapp.com/games/nfl-model-performance/. Accessed 8 Dec. 2023.

Keshtkar Langaroudi, Yamaghani, M. (2019). Sports result prediction based on machine learning and computational intelligence approaches: A survey. *Journal of Advances in Computer Engineering and Technology*, 5(1), 27–36.

Mortlock, D., Sanderson, D., &amp; Colombani, L. (2021, July 22). *How investment is changing sports.* Bain. https://www.bain.com/insights/how-investment-is-changing-sports/

Rookie Road, Inc.(2022, December 14), "What are Rushing Yards?", https://www.rookieroad.com/football/what-are-rushing-yards-3332685/

Warner, J. (2010). Predicting margin of victory in nfl games: Machine learning vs. the Las Vegas line. Published on: Dec, p. 17.